MONTHLY NOTICES

OF THE

ROYAL ASTRONOMICAL SOCIETY

Vol. 111 No. 5 1951

Published and Sold by the

ROYAL ASTRONOMICAL SOCIETY
BURLINGTON HOUSE
LONDON, W. 1

Price Nine Shillings

NOTICE TO AUTHORS

1. *Communications.*—Papers must be communicated to the Society by a Fellow. They should be accompanied by a summary at the *beginning* of the paper conveying briefly the content of the paper, and drawing attention to important new information and to the main conclusions. The summary should be intelligible in itself, without reference to the paper, to a reader with some knowledge of the subject; it should not normally exceed 200 words in length. **Authors are requested to submit MSS. in duplicate. These should be typed using double spacing and leaving a margin of not less than one inch on the left-hand side. Corrections to the MSS. should be made in the text and not in the margin.** Unless a paper reaches the Secretaries more than seven days before a Council meeting it will not normally be considered at that meeting. By Council decision, MSS. of accepted papers are retained by the Society for one year after publication; unless their return is then requested by the author, they are destroyed.

2. *Presentation.*—Authors are allowed considerable latitude, but they are requested to follow the general style and arrangement of *Monthly Notices*. References to literature should be given in the standard form, including a date, for printing either as footnotes or in a numbered list at the end of the paper. Each reference should give the **name and initials** of the author cited, irrespective of the occurrence of the name in the text (some latitude being permissible, however, in the case of an author referring to his own work). The following examples indicate the style of reference appropriate for a paper and a book, respectively:—

A. Corlin, *Zeits. f. Astrophys.*, **15**, 239, 1938.
A. S. Eddington, *Internal Constitution of the Stars*, Cambridge, p. 182, Table 24, 1926.

3. *Notation.*—Authors should conform closely to the recommendations of Commission 3 of the International Astronomical Union (*Trans. I.A.U.*, Vol. VI, p. 345, 1938). Council has decided to adopt the I.A.U. 4-letter abbreviations for constellations where contraction is desirable (Vol. IV, p. 221, 1932).

4. *Diagrams.*—**These should be drawn about twice the size required in print and prepared for direct photographic reproduction except for the lettering which should be inserted in pencil. Legends should be given in the manuscript indicating where in the text the figure should appear.** Blocks are retained by the Society for 10 years; unless the author requires them before the end of this period they are then destroyed.

5. *Tables.*—**These should be arranged so that they can be printed upright on the page.**

6. *Proofs.*—Costs of alteration exceeding 5 per cent of composition must be borne by the author. Fellows are warned that such costs have risen sharply in recent years, and it is in their own and the Society's interests to seek the maximum conciseness and simplification of symbols and equations consistent with clarity.

7. *Revised Manuscripts.*—When papers are submitted in revised form it is especially requested that they be accompanied by the original MS.

Reading of Papers at Meetings

8. When submitting papers authors are requested to indicate whether they will be willing and able to read the paper at the next or some subsequent meeting, and approximately how long they would like to be allotted for speaking.

9. Postcards giving the programme of each meeting are issued some days before the meeting concerned. Fellows wishing to receive such cards whether for Ordinary Meetings or for the Geophysical Discussions or both should notify the Assistant Secretary.

MONTHLY NOTICES

OF THE

ROYAL ASTRONOMICAL SOCIETY

Vol. 111 No. 5

ADDITIONAL MEETING OF 1951 JULY 25

in the Washington Singer Laboratories of the
University College of the South West, Exeter

Professor H. Dingle, President, in the Chair

The President announced that this Additional Meeting was the fourth to take place outside the rooms of the Society in London. He expressed the appreciation of the Society to the Principal and Council of the University College of the South West, through whose courtesy the meetings were being held in the Washington Singer Laboratories. The thanks of the Society were also accorded to the Director of the Norman Lockyer Observatory, Sidmouth, whose cooperation in the organization of the meeting was greatly appreciated.

One hundred and seventy-one presents were announced as having been received since the last meeting, including:—

G. de Vaucouleurs, *Physique de la Planète Mars* (presented by the author); and
M. Waldmeier, *Die Sonnenkorona* (presented by the author).

A number of Fellows signed the Admission Book. Papers were read by Dr W. H. Ramsey, Dr R. A. Lyttleton, Professor G. C. McVittie and Mr H. W. Newton.

Other meetings were held in connection with the Society's visit to Exeter. On the evening of July 23, a reception was given by the Principal and Council of the University College. An Astronomical Colloquium was held on July 24, and, in the evening, a Public Lecture, entitled New Astronomical Telescopes, was given by the Astronomer Royal; this lecture was illustrated by a film, "The Story of Palomar", lent by the American Embassy. A Geophysical Discussion on Geomagnetic Storms was held on July 26, and on the evening of the same day the Mayor of Exeter held a Civic Reception in the Guildhall for the Society. Two visits to the Norman Lockyer Observatory at Sidmouth were arranged, and also an excursion to places of archaeological interest on Dartmoor.

MEETING OF 1951 OCTOBER 12

Professor H. Dingle, President, in the Chair

The election by the Council of the following Fellows was duly confirmed:—

Rev. John Graham, M.A., B.Sc., Lecropt Manse, Bridge of Allan, Stirlingshire (proposed by W. M. Smart);

Don Tocher, A.B., Bacon Hall, University of California, Berkeley 4, California, U.S.A. (proposed by R. Stoneley); and

Wang Shdu-Kuang, University of London Observatory, Mill Hill Park, London, N.W.7 (proposed by G. R. Burbidge).

The election by the Council of the following Junior Member was duly confirmed:—

Roger John Tayler, Clare College, Cambridge (proposed by R. H. Garstang)

One hundred and fifty-four presents were announced as having been received since the last meeting, including:—

Atlas Stellarum Variabilium, Series IX, 1941 (presented by the Vatican Observatory);

Problems of Cosmical Aerodynamics (presented by the International Astronomical Union);

The Geomagnetic Field, its description and analysis, 1947, Publication No. 580 of the Carnegie Institution (presented by the Carnegie Institution of Washington); and

Astronomical Navigation Tables, volumes A–Q (presented by D. H. Sadler).

ON THE CONSTITUTIONS OF THE MAJOR PLANETS

W. H. Ramsey

(Received 1951 May 10)

Summary

The internal constitutions of the major planets are investigated on the basis of the atomic theory of solids. The starting-point of the investigation is the theoretical pressure–density relationship for solid hydrogen at absolute zero temperature; the internal temperatures of the planets are shown to be too low to influence the densities appreciably. It is shown from the empirical mean densities that the most massive planets, Jupiter and Saturn, are predominantly hydrogen. The hydrogen content of Jupiter is about 80 per cent by mass and that of Saturn about 60 per cent by mass. The proportions of hydrogen in these planets are comparable with that in solar material. The internal density distributions in Jupiter and Saturn have been estimated on the assumption of chemical homogeneity. The proportion of heavier elements has been chosen for each planet to agree with its empirical mean density. The computed moments of inertia of the planets are close to, but definitely larger than, the empirical values. This probably means that the heavier elements are not distributed uniformly in the planets but are to some extent centrally condensed. Uranus and Neptune are about ten times as dense as hydrogen planets of the same mass, and their mean atomic weights are about four; these planets are presumably composed mainly of water, methane and ammonia. The cosmogonical significance of these results is briefly discussed.

Introduction.—This investigation into the structures of the major planets has been prompted by recent theories of the internal constitution of the Earth. Seismic observations indicate a discontinuous change in structure at about halfway to the centre of the Earth, which is interpreted as the boundary of a central core. Until recently it has been taken for granted that the core is iron, probably alloyed with a small percentage of nickel. The less dense material of the surrounding mantle is mainly olivine, a mixture of magnesium and iron orthosilicates. The formation of an iron core is attributed to the immiscibility of molten iron and silicates; the iron is presumed to have separated out under gravity in the same manner as in an iron furnace. In the new theories the deep interior of the Earth is assumed to be chemically homogeneous, and the existence of a core is attributed to the enormous hydrostatic pressures at great depths. The pressure at the boundary of the core is supposed partially to destroy the molecular structure of the material, so that it becomes dense and metallic. The new view is consistent with all geophysical and astronomical data. The nature of the core is still open to debate, but the writer considers that the balance of the evidence is in favour of a chemically homogeneous Earth.

The challenge to the hypothesis of an iron core in the Earth has far-reaching implications, not only in geophysics but in planetary astronomy. Speculations on the internal structures of the major planets have followed the same general lines as those on the Earth. The major planets too are usually supposed to consist of a number of layers of different densities which have separated under gravity during the early history of the Solar System (*cf.* a recent "Report on the Progress

of Astronomy" by Wildt (1)). It is now in doubt if such chemical separations on a planetary scale could have occurred, even under the most favourable circumstances. Also, no allowance has been made for the drastic changes in molecular structure caused by the high pressures in the planets. Thus, if the new theories of the Earth be accepted, the existing models of the major planets must necessarily be revised. Such a revision will be attempted in the present paper.

Recent developments in geophysics.—Before 1936 every discontinuity in the Earth's internal structure was attributed to a change in chemical composition. In that year Bernal and Jeffreys suggested that the "20° discontinuity", which is at a depth of about 400 km., may be associated with the formation of a high-pressure modification of olivine. Jeffreys (2) has shown that this discontinuity is absent in the Moon, which indicates that it is a consequence of the increasing pressure rather than the appearance of a new chemical material. Arguing by analogy from the known properties of a related compound, magnesium germanate, Bernal (3) has suggested a possible crystal structure for the high-pressure modification of olivine. The estimated difference in density between the two phases is about nine per cent, and this is consistent with the geophysical evidence. This explanation of the "20° discontinuity" greatly simplified the picture of the Earth. It soon became widely accepted that the Earth is, to a good approximation, chemically homogeneous between the base of the crustal layers (depth 33 km.) and the boundary of the central core (depth 2900 km.). The core discontinuity alone was interpreted in terms of a change in chemical composition.

In 1941 Kuhn and Rittmann (4) published an important paper in which they contended that the discontinuous behaviour of seismic waves at the boundary of the Earth's core is not associated with a sudden change in the chemical composition. These authors adopted a novel approach to the problem of the Earth's internal constitution. Instead of starting from the known geophysical facts, which is the usual approach, they attempt to follow the Earth's evolution on the assumption that it was formed from material ejected from the Sun. Their first conclusion is that an iron core could not have been formed in the Earth. They estimate that, even under the most favourable circumstances, the chemical separation would require a time much greater than the age of the Earth. Their second conclusion is that the deep interior of the Earth is, even to-day, of solar composition. The authors claim that, while the light volatile materials have escaped from the outer layers of the Earth, diffusion is too slow a process to have removed an appreciable fraction of the hydrogen from the core. Kuhn and Rittmann attempt to reconcile this model of the Earth with the seismic data. Hitherto the discontinuous behaviour of seismic waves at the boundary of the core has been associated with a jump in density. Kuhn and Rittmann reject this interpretation on the grounds that it is inconsistent with their main thesis, namely, that the deep interior of the Earth is chemically homogeneous and of solar composition. Instead they ascribe the apparently discontinuous behaviour to a very rapid decrease in the viscosity. They assume that just above the core boundary the material is sufficiently viscous to transmit seismic vibrations with the characteristics of a solid, and that just below the boundary the characteristics are typical of a liquid. The core discontinuity is remarkably sharp; its reflecting power for seismic waves shows that the transition from mantle to core, if not strictly discontinuous, cannot be spread over more than three or four kilometres.

The explanation of the core advanced by Kuhn and Rittmann therefore assumes an extremely rapid decrease in the viscosity without an appreciable change in the chemical composition, and this assumption needs justification. Kronig, de Boer and Korringa (**5**) pointed out that a sudden decrease in viscosity at the boundary of the core would be understandable if the hydrogen in the solar material became metallic at the pressure at this depth, and to test this idea they made detailed calculations on the pressure–density relationship for hydrogen. They found that the critical pressure for the transition of pure hydrogen to the metallic phase is smaller than that at the boundary of the core, but only by a factor of two; the discrepancy could be due to the presence of the heavier elements in the solar material. This was the first attempt to associate the core discontinuity with the transition of a non-metal to a metallic phase, and it preceded the writer's work on the Earth by more than two years. This ingenious theory of Kuhn and Rittmann is untenable for the following reasons. The postulate of solar composition is inconsistent with the high density of the material in the deep interior of the Earth. Solar matter is predominantly hydrogen, and Kronig, de Boer and Korringa (**5**) have shown that the density of hydrogen at the pressures in the core is only about 1 g./cm.^{3}. A density in the neighbourhood of 10 g./cm.^{3} is necessary to explain the mean density and moment of inertia of the Earth. Thus the hydrogen and helium must have escaped from the interior of the Earth, if indeed these elements were ever present in large quantities. The conclusion that the volatile materials could not have escaped from the interior of the Earth is based on the assumption that the molecules must first come to the surface by the slow process of diffusion. But diffusion is not a controlling factor if the planet is cooling by convection, and Jeffreys (**6**) considers that there would be ample time for the gases from the interior to escape before convection ceased. The theory of Kuhn and Rittmann is also at variance with the low mean densities of the major planets. Since the interior of every planet is supposed to be of solar composition, the larger planets should be denser because of the greater compression. But the Earth is about eight times as dense as Saturn, though its mass is smaller by a factor of a hundred. It will be shown in this paper that only the largest planets in the Solar System, Jupiter and Saturn, are even approximately of solar composition.

The writer (**7, 8**) has recently developed a theory in which the Earth's core and mantle are assumed to be chemically identical. This theory is an extension of the work of Bernal and Jeffreys, and it is essentially different from the theory of Kuhn and Rittmann. The discontinuous behaviour of seismic waves at a depth of 2900 km. is interpreted in the usual way in terms of a jump in density. But the jump in density is attributed to a phase transition under pressure and not to the appearance of a new material such as iron. In the mantle olivine has an ionic structure; in the core it is in a dense metallic state. Phase transitions of this type have been discussed at length in other papers (**8**). A transition to a metallic phase is more drastic than an ordinary polymorphic transition such as that suggested by Bernal and Jeffreys to explain the "20° discontinuity". In ordinary phase transitions the atoms or molecules are merely rearranged into a different crystal pattern and the density changes by only a few per cent. But in a transition to a metallic phase the atoms or molecules themselves are partially destroyed and electrons are set free to wander through the crystal. The density increases by more than 50 per cent at the boundary of the Earth's core, and this

is consistent with the view that the material becomes metallic. The critical pressure for a transition to a metallic phase is usually very large by laboratory standards. A considerable energy is expended in releasing the electrons when the material becomes metallic, and this energy must be supplied by the hydrostatic pressure during the collapse to the denser phase. The critical pressure will therefore be high, and it has been shown in another paper (**8**) that it will usually be of the order of magnitude of a million atmospheres. The pressure at the boundary of the Earth's core is about 1·4 million atmospheres. The Earth also possesses a small inner core of radius about 1250 km.; this is attributed to a second phase transition which increases the number of "free" electrons. The writer's theory, together with the work of Bernal and Jeffreys, implies that the Earth is chemically homogeneous below the base of the crustal layers (depth 33 km.). The gradual increase in density with depth is due to the increase in compression, and the abrupt jumps in density are attributed to phase transitions under pressure. The pressure–density relationship for the material of the Earth can be computed directly from the internal density distribution which has been derived from the seismic data (**9**). This theory is compatible with all data on the Earth's internal structure, and it can account for several features which have not been previously explained (**8**). The mean densities of the terrestrial planets probably provide the strongest evidence in support of the theory. Jeffreys (**10**) has pointed out that the various terrestrial planets have different chemical compositions, assuming that the Earth's core and mantle are chemically distinct. The smaller planets have relatively smaller cores, which, on traditional views, means that they are deficient in iron. The smaller planets are unlikely to have retained (or captured) a higher percentage of the lighter elements, and so it would appear that the planets were originally of different chemical compositions. The new interpretation of the core removes this difficulty. The smaller planets have relatively smaller cores because they have smaller central pressures, and in the smallest planets the central pressure cannot support a core at all. It has been shown in another paper (**7**) that, on the basis of the new theory, the Earth and the terrestrial planets have the same chemical composition. This composition is essentially the same as that of meteorites, and it is consistent with solar origin.

These developments in geophysics have important repercussions on the theories of the internal constitutions of the major planets. The most widely accepted models of the major planets are those proposed by Wildt (**1, 11**). He assumes that the different components of the originally gaseous mixture separated under gravity in much the same way as an iron core is supposed to have formed in the Earth. The details of the separation processes have been filled in by speculations along geochemical lines. Each major planet is presumed to have a central core of terrestrial composition; the mean density of this core is usually put at 6 g./cm.^{3}, which is a little greater than the mean density of the Earth. Around this core a thick layer of "ice" is supposed to have formed, the chief constituents being water, carbon dioxide and other vapours which condense readily; the mean density of this layer is estimated to be about 1·5 g./cm.^{3}. Finally, the outermost layer is presumed to be condensed permanent gases in a state of high compression, the predominant constituent being solid hydrogen; the mean density of this layer is assumed to be 0·3 g./cm.^{3}. This model contains only two adjustable parameters, namely, the radii of the spheres of discontinuity. The thicknesses of the various layers can be determined for each planet from its

known mean density and moment of inertia. This model leads to unreasonable chemical compositions for the major planets. For example, Wildt finds that Saturn, which is intermediate in mass, contains relatively about three times as much free hydrogen as the other major planets. This conclusion, were it valid, would mean that the planets were originally of different chemical compositions. But the predictions of the model are unreliable as no allowance is made for the transitions to metallic phases which occur at the high pressures in the interiors of the planets. For example, it is assumed in the model that the density of a terrestrial core at the centre of a major planet is only 6 g./cm.^{3}. But the density of terrestrial material is already about 20 g./cm.^{3} at the pressure at the centre of the Earth, and it will be at least 40 g./cm.^{3} at the pressure at the centre of Jupiter. The densities used for the other layers are also too low. The heavy elements are therefore not so abundant in the major planets as the model suggests. As long ago as 1938 Wildt (**11**) and Kothari (**12**) independently suggested that hydrogen may be metallic in the interiors of the major planets. This suggestion has been confirmed by the detailed calculations of Kronig, de Boer and Korringa (**5**). These calculations form the basis of the present investigation, and it will be shown later in the paper that hydrogen is the predominant constituent of both Jupiter and Saturn. The feasibility of large-scale chemical separations under gravity was taken for granted when Wildt first proposed his models of the major planets. But meanwhile Kuhn and Rittmann have shown that geological time is too short to permit the formation of an iron core in the Earth. The time required to effect a chemical separation under gravity increases rapidly with the dimensions of the planet. It therefore seems probable that no appreciable chemical separation has occurred in the major planets since their formation. The major planets may therefore be chemically homogeneous. But, in the writer's opinion, the issue of chemical homogeneity cannot be settled by *a priori* considerations alone. The present state of the major planets depends critically on their mode of formation, and the origin of the Solar System is still an open question. Each major planet will possess a small central core of terrestrial composition if the planets were formed by an aggregation process. On the other hand, the planets could be chemically homogeneous if they were formed from material ejected from a star. It will be assumed in the present paper that the major planets are chemically homogeneous. The possibility of a dense terrestrial core will be examined in a later paper.

Empirical data.—The masses, mean diameters and mean densities of the major planets are given in Table I; the data for the Earth are also given for comparison. The values quoted and the estimated errors have been taken mainly from Russell, Dugan and Stewart (**13**), but Kuiper's (**14**) new determination of the mean diameter of Neptune has been adopted. The masses of the major planets are known accurately as all of these planets have satellites. The uncertainties in the mean densities arise almost entirely from the errors in the determinations of the diameters. The two more remote planets, Uranus and Neptune, have very small angular dimensions. The angular diameter of Neptune is only about two seconds of arc, as compared with fifty seconds for Jupiter at closest approach. Also, the more remote planets are less brilliantly illuminated. The mean diameters of Jupiter and Saturn are known to about one per cent, but that of Uranus is known only to about six per cent. Kuiper's (**14**) recent measurement of the diameter of Neptune involves a new

technique, and he considers that his determination is correct to about one per cent; the most reliable previous estimates are too large by more than ten per cent. In Table I the mean density appears to vary erratically from planet to planet. One would expect the mean density of a planet to be determined primarily by its mass, assuming that the planets were formed by a single event or process from a common original material. But Saturn, which is intermediate in mass, is only about half as dense as the other major planets. There are, however, two opposing factors to be considered. First, the greater compression in the larger planets tends to make the mean density increase with the mass. Second, the larger planets may have retained (or captured) a higher percentage of the lighter elements, which would tend to make the mean density decrease with the mass. The net result may be that the mean density is a complicated function of the mass. It will be seen later that hydrogen is relatively more abundant in Jupiter and Saturn than in the smaller planets, Uranus and Neptune.

TABLE I

Mean Densities of Major Planets

	Relative Mass	Mean Diameter (km.)	Mean Density (g./cm.^{3})
Earth	1	12,742	5·52
Jupiter	318·35 ± 0·01	139,800 ± 1,400	1·33 ± 0·04
Saturn	95·3 ± 0·1	115,100 ± 1,200	0·71 ± 0·02
Uranus	14·6 ± 0·1	51,000 ± 3,000	1·3 ± 0·2
Neptune	17·3 ± 0·1	44,600 ± 400	2·22 ± 0·07

TABLE II

Ellipticities and Moments of Inertia

	e^{-1}	m^{-1}	e/m	I/MR^2
Earth	297·3	289·8	0·975	0·334
Jupiter	15·4 ± 0·2	11·2 ± 0·3	0·73 ± 0·02	0·25
Saturn	9·5 ± 0·2	6·2 ± 0·2	0·65 ± 0·02	0·22
Uranus	18·0 ± 0·6	12·5 ± 2	0·69 ± 0·11	0·24
Neptune	49·0 ± 2·5	38·0 ± 1	0·78 ± 0·04	0·27

The mean density is not the only indication of a planet's internal constitution. Additional information can be obtained from the polar flattening caused by its rotation. In Table II the ellipticity e is the difference between the equatorial and polar diameters, expressed as a fraction of the equatorial diameter. The ellipticity can be estimated directly from visual measurements or it can be calculated from the precession of the orbit of a satellite. The latter method is the more accurate, and all the values quoted in Table II have been determined in this way. The ellipticities given for Jupiter and Saturn are due to H. Struve (*cf.* reference (**13**)) and that for Uranus to Lvoff (**15**). Lvoff's calculations on Neptune have also been used, but they have been modified as required by Kuiper's new determination of the planet's diameter. Neptune's axial rotation cannot be detected directly as there is no visible surface marking. The rate of rotation is measured indirectly by means of the Doppler effect, which gives the velocity on the planet's equator. The period of rotation can be estimated only in terms of the diameter, and the older measurements of the diameter correspond to a period of about 15·5 hours. Using Kuiper's new determination of the diameter, the estimated period of axial rotation is reduced to 13·8 hours. The

ellipticity of Neptune given in Table II follows directly from this period of rotation and Kuiper's estimate of the diameter (*cf.* Lvoff (**15**)). From the point of view of a planet's internal structure, the interesting quantity is not the ellipticity e itself but the ratio (e/m), where m is the ratio of the centrifugal and gravitational forces on the planet's equator. The ratio (e/m) is smaller for planets whose mass is more concentrated towards the centre; it is 1·25 for a planet of uniform density, and it is 0·5 if the mass of the planet is concentrated in a small central core. Among the major planets (e/m) ranges from 0·65 to 0·78 and so all of these planets have strong central condensations of mass. They are even more markedly condensed than the Earth, which itself has a dense central core. The extent of the central condensation is indicated quantitatively by the ratio (I/MR^2), where M, R and I are the planet's mass, radius and moment of inertia respectively. The magnitude of (I/MR^2) is determined in terms of the ratio (e/m) by the well-known Radau–Darwin formula (*cf.* reference (**16**)):

$$I/MR^2 = \tfrac{2}{3}\{1 - \tfrac{2}{5}\sqrt{(1+\eta)}\}, \qquad (1)$$

where the parameter η is defined by the relation

$$\eta = 5m/2e - 2. \qquad (2)$$

The values of (I/MR^2) calculated in this way are given in the last column of Table II. For the major planets the ratio (I/MR^2) ranges from 0·22 to 0·27, as compared with 0·4 for uniform density and 0·334 for the Earth. The central condensations are therefore very marked indeed. Saturn is probably the most centrally condensed of the major planets, and Neptune is probably the least. Equations (1) and (2) are not exact; second and higher powers of the ellipticity have been ignored, and a certain function has been put equal to unity (*cf.* reference (**16**)). A more accurate treatment requires a knowledge of the internal density distribution, and this is not available. Equations (1) and (2) are valid approximations for the major planets, but the calculated values of (I/MR^2) may be in error by several per cent.

The data in Tables I and II can be combined on reasonable assumptions to set definite limits to the surface and central densities of the planets. An upper limit to the surface density ρ_s is set by the following argument due to Jeffreys (**17**). It is assumed that the density at the surface is less than that at any internal point. The planet's moment of inertia I is greater than it would be if the whole volume had the surface density ρ_s, the mean density being maintained at its actual value by a mass point at the centre; that is,

$$I > (8\pi/15)\rho_s R^5,$$

from which it follows that

$$\rho_s/\bar{\rho} < 2{\cdot}5 I/MR^2, \qquad (3)$$

$\bar{\rho}$ being the mean density of the planet. The same type of argument can be used to set a lower limit to the central density ρ_c. It is assumed that the central density is the largest in the planet. The moment of inertia I is larger than it would be if the planet consisted of a central core of radius a and density ρ_c, surrounded by an envelope of negligible mass; that is,

$$I > (8\pi/15)\rho_c a^5, \qquad \rho_c a^3 = \bar{\rho} R^3,$$

from which it follows that

$$\rho_c/\bar{\rho} > \{2{\cdot}5 I/MR^2\}^{-3/2}. \qquad (4)$$

The limits set to ρ_s and ρ_c by equations (3) and (4) are given in Table III. The upper limits to the surface densities are surprisingly small, especially for Saturn.

TABLE III

Limits to Surface and Central Densities

(g./cm.^{3})

		Earth	Jupiter	Saturn	Uranus	Neptune
ρ_s	<	4·6	0·83	0·39	0·8	1·5
ρ_c	>	7·3	2·7	1·7	2·8	4·0

Moreover, it must be remembered that these limits have been derived on an extreme hypothesis. The actual surface densities will be much smaller, and throughout considerable fractions of the volumes the densities must be less than these limits. Hydrogen and helium are the only solids which have densities as small as 0·39 g./cm.^{3}, and the outer layers of Saturn must consist mainly of these elements. The presence of an atmosphere is of little importance so far as the bulk properties of a major planet are concerned. Peek (**18**) has pointed out that all gases in the atmospheres of the major planets are highly compressed under their own weight. The density of a perfect gas would exceed that of its solid at a depth below the visible surface which is less than one per cent of the radius. Thus the perfect gas laws are applicable only in a thin surface layer whose contributions to the mass and volume of the planet are negligible. The remainder of the planet is probably solid. But even if part of it is gaseous, the equation of state of a gas at these high pressures approximates to that of the solid. It will be assumed in this paper that the planets are solid throughout.

Hydrogen planets.—If a planet is sufficiently massive to have retained (or captured) even the lightest volatile compounds, one would expect it to be approximately of solar composition. As over 90 per cent of all atoms in the Sun's atmosphere are hydrogen, the other elements in solar material may be neglected as a first approximation. In this section of the paper the properties of planets composed of hydrogen will be investigated, and the mean density will be computed as a function of the mass of the planet. A comparison with the data in Table I will throw some light on the chemical constitutions of the major planets.

TABLE IV

Pressure–Density Relationship for Hydrogen

Molecular Phase		Metallic Phase	
Pressure 10^{12} dynes/cm.^{2}	Density g./cm.^{3}	Pressure 10^{12} dynes/cm.^{2}	Density g./cm.^{3}
0	0·089	0·8	0·77
0·025	0·167	1·0	0·81
0·05	0·195	2·0	0·98
0·1	0·225	3·0	1·12
0·2	0·261	4·0	1·25
0·3	0·284	5·0	1·36
0·4	0·302	10·0	1·84
0·5	0·316	15·0	2·21
0·6	0·329	20·0	2·53
0·7	0·340	30·0	3·07
0·8	0·350	40·0	3·53

The pressure–density relationship for solid hydrogen at absolute zero temperature is given in Table IV. The critical pressure* for the transition

* One atmosphere is $1·01 \times 10^6$ dynes/cm.^{2}.

of hydrogen to a metallic phase is $0{\cdot}8 \times 10^{12}$ dynes/cm.^{2}. Below the critical pressure the densities refer to the molecular phase and at higher pressures to the metallic phase. This pressure–density relationship has been computed numerically on the basis of the atomic theory of solids (**19**). The calculations of Kronig, de Boer and Korringa (**5**) have been used for both the molecular and metallic phases. The writer (**19**) has, however, refined the calculations of Kronig *et al.* in one respect; the zero-point energies of the molecules in molecular hydrogen and of the protons in metallic hydrogen have been taken into account by an approximate procedure due to Debye, *cf.* reference (**20**). The zero-point energy always decreases the density at a given pressure. It exerts a strong influence at low pressures; if it is neglected, the calculated density of solid hydrogen at zero pressure is about 60 per cent larger than the empirical value. The correction to the density is much less at higher pressures and, over most of the pressure range which is of interest, the correction is about 7 per cent. The calculated critical pressure for the formation of metallic hydrogen is not altered appreciably by the inclusion of the zero-point energies. Kronig *et al.* estimated the critical pressure to be $0{\cdot}7 \times 10^{12}$ dynes/cm.^{2} and, at the critical pressure, the densities of the two phases were found to be 0·4 and 0·8 g./cm.^{3}. According to the present calculations the critical pressure is $0{\cdot}8 \times 10^{12}$ dynes/cm.^{2}, and at this pressure the densities of the two phases are 0·35 and 0·77 g./cm.^{3}. As regards the reliability of the data in Table IV, the uncertainties in the densities of the metallic phase are too small to be of any consequence. From the point of view of atomic theory, metallic hydrogen is the simplest of all solids. It is simpler than the other alkali metals in that the ion core is a proton which, rigorously, has a Coulomb field. Complex ions introduce difficulties which at present limit the accuracy of calculations on solids. These difficulties do not arise with metallic hydrogen, and the pressure–density relationship can be computed using existing techniques to an accuracy which is more than adequate for the present purposes. The calculated densities of solid molecular hydrogen are less reliable, and an error of five per cent would not be surprising. The critical pressure is the most uncertain quantity in Table IV and it may be in error by ten, or even twenty, per cent. It will be seen later that these uncertainties are not an important limitation for the problem of the planets.

As the density ρ of solid hydrogen is known as a function of the pressure p, the bulk properties of planets composed of hydrogen can be calculated on the assumption of hydrostatic equilibrium. The equation to be solved is

$$dp/dr = -\rho g, \tag{5}$$

where r is the distance from the centre of the planet and where g is the local acceleration due to gravity. This equation has been solved for a number of arbitrarily chosen central pressures, using step-by-step numerical integration. The solution of equation (5) gives the density as a function of the distance from the centre of the planet. This internal density distribution determines uniquely the mass M, the mean density $\bar{\rho}$ and all the other mechanical properties of the planet. The results of the calculations are presented in Fig. 1, which shows the dependence of the calculated mean density on the mass of the planet. The observed mean densities of the four major planets are also shown for comparison. The central pressure in a small planet is too low to support a core of metallic hydrogen. A core is possible only if the mass of the planet exceeds $88 M_g$,

M_E being the mass of the Earth. A planet whose mass lies between $88M_E$ and $95M_E$ can exist in equilibrium with or without a core. This is a consequence of the fact that the density of hydrogen jumps by more than 50 per cent at the critical pressure. The equilibrium configurations which have the smallest cores are unstable and so would not occur in nature; this leads to the discontinuity in the curve in Fig. 1. This phenomenon was encountered in previous calculations on the terrestrial planets (**7**), and a full explanation has been given by Lighthill and the writer (**21, 22**). A planet always has a core of metallic hydrogen if its mass exceeds $95M_E$, and the fraction of the total mass which is contained in the core increases rapidly with the mass of the planet. As is to be expected, all the major planets are denser than they would be if composed entirely of solid hydrogen. But the largest planets, Jupiter and Saturn, are only about twice as dense as hydrogen planets of the same mass. This difference in mean density could be removed by increasing the density function $\rho(p)$ by about 20 per cent, and solar material may be denser than pure hydrogen by about this amount.

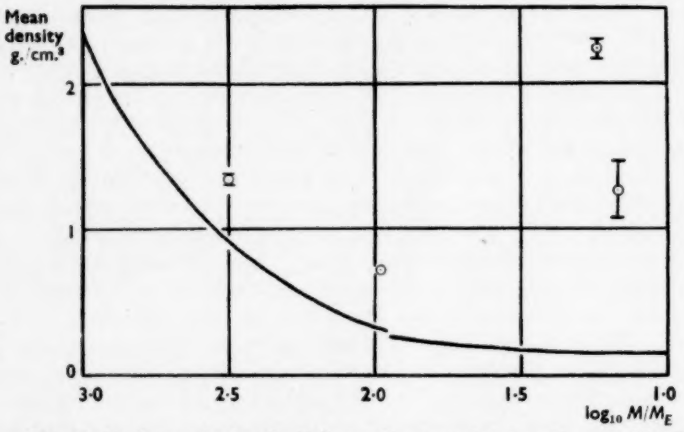


FIG. 1.—*The calculated mean density of a hydrogen planet as a function of its mass M; M_E is the mass of the Earth. The empirical data for the major planets are also shown (cf. Table I).*

Jupiter and Saturn therefore consist of hydrogen to about 80 per cent by mass and are thus much closer to solar composition than has been hitherto supposed; Wildt (**11**) put the hydrogen content of Jupiter as low as 10 per cent by mass. On the other hand, Uranus and Neptune are more than ten times as dense as hydrogen planets of the same mass. A fourfold increase in the density function $\rho(p)$ would be necessary to remove this discrepancy, and so the pressure–density relationship for solid hydrogen is not a proper starting-point for an investigation into the internal structures of these planets. The mean atomic weight in Jupiter and Saturn is little more than unity, but in Uranus and Neptune it is probably in the neighbourhood of four. Thus, from the point of view of chemical composition the major planets fall into two groups rather than one. The remainder of this paper deals mainly with the planets of the first group, Jupiter and Saturn, which are predominantly hydrogen. The calculations on hydrogen planets will be refined so as to include the relatively small proportions of heavier elements necessary to give agreement with the observed mean densities. A quantitative

treatment of the bulk properties of the planets of the second group, Uranus and Neptune, must await the results of calculations on the pressure–density relationships of a number of solids. The cosmogonical significance of this grouping of the planets according to chemical composition will be discussed later in the paper.

Rotating planets of solar composition.—The internal density distributions in Jupiter and Saturn will now be estimated on the assumption of chemical homogeneity. The calculations on hydrogen planets must be modified in two respects. First, allowance must be made for the presence of the heavier elements. Second, the centrifugal force arising from the planet's axial rotation must be taken into account in the hydrostatic equation (5). The centrifugal force effectively decreases gravity and so also the pressures in the interior of the planet; it therefore tends to reduce the mean density. The centrifugal force is relatively more important in Jupiter and Saturn than in the Earth. The ratio m of the centrifugal and gravitational forces on the equator is given in Table II for the different planets. This ratio is only 0·3 per cent for the Earth but it is 9 per cent for Jupiter and 16 per cent for Saturn. The polar flattening caused by the axial rotation destroys the spherical symmetry of the planet. The hydrostatic equation therefore involves two independent variables, the distance from the centre of the planet and the angle of latitude. An exact treatment of this equation would be cumbersome, but an approximate solution can be obtained by a method often employed in problems concerned with the figure of the Earth, *cf.* reference (**23**). It will be assumed that the centrifugal force is so much smaller than gravity that the square of the ratio m may be neglected in comparison with unity. One of the independent variables can be eliminated from the hydrostatic equation by the following transformation. Any given pressure p defines in the planet an oblate surface which is approximately a spheroid. Let r be the radius of the sphere which contains exactly the same volume as this surface. The parameter r is introduced as the new independent variable, and it takes the place of the distance from the centre of the planet in the simple hydrostatic equation (5). It may easily be established by the methods described by Jeffreys and Jeffreys (**23**) that the equation for hydrostatic equilibrium, neglecting terms of the order of m^2, is

$$dp/dr = -\rho\{g(r) - \tfrac{2}{3}\omega^2 r\}, \qquad (6)$$

where ω is the angular velocity of axial rotation and where $g(r)$ is the acceleration due to gravity:

$$g(r) = (4\pi G/r^2)\int_0^r \rho(x)x^2\,dx,$$

G being the constant of gravitation. Equation (6) replaces the simple hydrostatic equation (5) and it is to be solved as though the planet had spherical symmetry; the oblateness is allowed for in the definition of the parameter r. Equation (6) differs from the simple equation (5) only by the term $\frac{2}{3}\omega^2 r$ in the curly brackets. The quantity $\frac{2}{3}\omega^2 r$ is the average value of the radial component of the centrifugal acceleration over a central sphere of radius r. The inclusion of the centrifugal force reduces the calculated mean densities of planets similar to Jupiter and Saturn by two and five per cent respectively. These corrections are smaller than would be expected from the ratios of the centrifugal and gravitational forces at the surface, but the centrifugal force is relatively less important in the

interiors of the planets. The uncertainties in the calculated mean densities due to the neglect of second and higher powers of the ratio m in equation (6) are only of the order of 0·1 per cent.

Next to hydrogen, helium is the most abundant element cosmically. The mass of helium in solar material is about ten times as great as the total mass of all heavier elements. Hydrogen and helium together account for over 99 per cent of the total mass, and so the elements heavier than helium have little influence on the pressure–density relationship. It will be assumed in the following treatment that Jupiter and Saturn are composed entirely of hydrogen and helium, and the proportion of helium to hydrogen will be determined for each planet from its observed mean density. Heavier elements are undoubtedly present in Jupiter and Saturn; methane (CH_4) and ammonia (NH_3) have been detected spectroscopically in the atmospheres of both planets. But unless the heavier elements are many times as abundant as in solar material, they will make only a minor contribution to the bulk properties of the planets. Moreover, since the proportion of helium is determined from the empirical mean density, the neglect of the heavier elements is compensated by a corresponding increase in the helium content. Let ρ_H and ρ_{He} be the densities of hydrogen and helium respectively at pressure p. It is convenient to express ρ_{He} as a multiple of ρ_H:

$$\rho_{He}=\alpha\rho_H. \tag{7}$$

Consider a mixture which contains n times as many hydrogen as helium atoms; the proportions of hydrogen and helium by mass are in the ratio of n to 4. It will be assumed that the volume occupied by the mixture is equal to the sum of the volumes occupied by the components separately. The density of the mixture at pressure p is therefore

$$\rho=\{(n+4)/(n+4\alpha^{-1})\}\rho_H. \tag{8}$$

If the ratio α is known, this relation expresses the equation of state of the mixture in terms of the data for hydrogen in Table IV. Detailed calculations on the pressure–density relationship for helium are not available, and so the ratio α can only be estimated by an approximate method. At very high pressures the density of a solid is determined mainly by the repulsive interactions between the constituent atoms or ions (**8**). Lennard-Jones (**24**) has derived an expression for the repulsion between two helium atoms, and this has been used to estimate the density of ordinary solid helium as a function of the pressure. It is found that solid helium is about four times as dense as solid molecular hydrogen at the high pressures important in the planets. In other words, at high pressures the ordinary solid phases of hydrogen and helium occupy approximately the same volume per atom. The value of α is therefore four in the non-metallic mantles of the planets,

$$\alpha=4 \qquad \text{(mantle)}, \tag{9}$$

and the equation of state is

$$\rho=\{(n+4)/(n+1)\}\rho_H \qquad \text{(mantle)}. \tag{10}$$

An estimate of the ratio α for the metallic cores can be made as follows. The electrostatic interactions among the electrons and the nuclei are of no importance at extremely high pressures such as exist in the white dwarf stars (**19**). The electrons constitute a degenerate Fermi gas, and the density at a given pressure is proportional to the ratio of atomic weight to atomic number. This ratio is two for helium and unity for hydrogen. Metallic helium is therefore only twice

as dense as metallic hydrogen at the pressures prevailing in the interiors of the white dwarf stars. The electrostatic interactions among the electrons and the nuclei are important at the lower pressures in the planets, and the ratio α is accordingly greater than two. The correction to α due to the electrostatic interactions can be evaluated approximately by Thomas–Fermi calculations. Such calculations have recently been made by Scholte (**25**) and he finds that helium is about 2·5 times as dense as hydrogen at the pressures in the metallic cores of the planets:

$$\alpha = 2{\cdot}5 \quad \text{(core)}. \tag{11}$$

The reliability of Thomas–Fermi calculations at planetary pressures has been discussed in another paper (**19**). Only the most important electrostatic interactions are taken into account in these calculations and the remaining interactions would probably raise the calculated value of α. But the increase is unlikely to be large; the neglect of all electrostatic interactions reduces the calculated ratio (11) by only 20 per cent. From equations (8) and (11) the density of the material in the metallic cores is

$$\rho = \{(n+4)/(n+1{\cdot}6)\}\rho_H \quad \text{(core)}, \tag{12}$$

ρ_H being the density of metallic hydrogen at the pressure concerned. For a given value of n, equations (10) and (12) give the pressure–density relationship for the mixture in terms of the data in Table IV for hydrogen; equation (10) is to be used at pressures below the critical pressure for the production of metallic hydrogen and equation (12) at higher pressures. The hydrostatic equation (6) has been solved numerically for a series of arbitrary central pressures and a number of values of n. For a given central pressure and a given value of n, the solution of equation (6) determines the mass, radius and internal density distribution of the planet. The central pressure and the value of n have been adjusted by trial and error until the planet has the desired mass and radius. The central pressure $31{\cdot}5 \times 10^{12}$ dynes/cm.^{2} and n equal to 13·0 correspond to the mass and radius of Jupiter; the central pressure $6{\cdot}00 \times 10^{12}$ dynes/cm.^{2} and n equal to 6·4 correspond to the mass and radius of Saturn. The calculated internal density distributions of these planets are presented in Tables V and VI. In these tables r is the parameter appearing in equation (6), p and ρ are the corresponding pressure and density, and $M(r)$ is the mass contained within the surface defined by the parameter r:

$$M(r) = 4\pi \int_0^r \rho(x) x^2 \, dx.$$

According to the present calculations the number of hydrogen atoms per atom of helium is 13·0 in Jupiter and 6·4 in Saturn. The hydrogen content of Jupiter is therefore 76 per cent by mass and of Saturn 62 per cent by mass. Thus Jupiter and Saturn contain much more hydrogen than was apparent from Wildt's models. Moreover, the hydrogen contents of the planets are probably underestimated in the present calculations. The proportion of helium to hydrogen has been estimated from the empirical mean density on the assumption of chemical homogeneity. At the instigation of the writer, Mr B. Miles is at present making calculations on Jupiter on the assumption that the heavier elements are concentrated around the centre of the planet. It is found that this central concentration reduces the proportion of helium necessary to account for the mean density of Jupiter from 24 to 21 per cent by mass; the calculated hydrogen content is

therefore raised to 79 per cent. The hydrogen content is further increased to 85 per cent by mass if the helium is replaced by heavy elements such as silicon and iron. The hydrogen content of Saturn may likewise be underestimated. Thus the present calculations set lower limits to the proportions of hydrogen in the planets. The actual hydrogen contents depend on the relative abundances of the elements heavier than hydrogen and on their spatial distributions within the planets.

TABLE V

Estimated Density Distribution in Jupiter

($n=13{\cdot}0$)

r 10^9 cm.	p 10^{12} dynes/cm.^{2}	ρ g./cm.^{3}	$M(r)$ 10^{27} g.
0	31·5	3·66	0
1·0	29·7	3·56	15·0
1·5	27·6	3·43	50·2
2·0	24·9	3·27	115
2·5	21·6	3·05	215
3·0	18·1	2·80	355
3·5	14·5	2·53	532
4·0	11·0	2·23	743
4·5	7·86	1·92	979
5·0	5·09	1·60	1228
5·5	2·82	1·28	1475
6·0	1·09	0·96	1709
6·10	0·80	0·89	1751
6·10	0·80	0·42	1751
6·3	0·56	0·39	1790
6·5	0·35	0·36	1827
6·7	0·17	0·30	1863
6·9	0·04	0·22	1894
6·99	0	0·11	1902

TABLE VI

Estimated Density Distribution in Saturn

($n=6{\cdot}4$)

r 10^9 cm.	p 10^{12} dynes/cm.^{2}	ρ g./cm.^{3}	$M(r)$ 10^{27} g.
0	6·00	1·91	0
1·0	5·53	1·85	7·8
1·5	4·97	1·77	26·0
2·0	4·25	1·66	59·1
2·5	3·43	1·53	110
3·0	2·56	1·38	179
3·5	1·71	1·21	264
4·0	0·92	1·03	363
4·08	0·80	1·00	380
4·08	0·80	0·49	380
4·5	0·53	0·45	426
5·0	0·26	0·39	486
5·2	0·17	0·35	511
5·4	0·095	0·31	535
5·6	0·032	0·24	556
5·755	0	0·125	569

In Jupiter the metallic core contains 92 per cent of the total mass and in Saturn 67 per cent of the total mass. The bulk properties of the planets are therefore determined mainly by the pressure–density relationship for metallic hydrogen, which is known accurately. The pressure–density relationship for

solid molecular hydrogen and the critical pressure for the production of metallic hydrogen need not be known with great accuracy for the present calculations, and these are the most uncertain quantities in Table IV. The critical pressure given in Table IV may be in error by 20 per cent and the densities of solid molecular hydrogen by 5 per cent. The total radii of the planetary models in Tables V and VI are decreased by about 1 per cent if the critical pressure adopted for hydrogen is reduced by 20 per cent, the masses and compositions of the planets being assumed unchanged. The total radii of these model planets are also decreased by about 1 per cent if the estimated densities of solid molecular hydrogen are raised by 5 per cent. These errors are not independent but tend to cancel one another; the critical pressure is automatically increased by about 10 per cent if the densities of the molecular phase are raised by 5 per cent. Thus the uncertainties in the data for hydrogen in Table IV introduce an uncertainty of less than 1 per cent in the computed radii of the planetary models under consideration. Equation (8) has been used to estimate the increase in density when one helium atom is added for every n atoms of hydrogen. This equation involves the ratio α of the densities of helium and hydrogen at the pressure concerned. As detailed calculations on helium are not available, the ratio α is necessarily uncertain and the values (9) and (11) adopted may be in error by 10, or even 20, per cent. But the proportion of helium is small, and the density for a given value of n is uncertain by only 1 or 2 per cent. As an error of a few per cent can be tolerated in the densities in the non-metallic mantles of the planets, equation (10) gives the pressure–density relationship of the mixture with sufficient accuracy. An error in the adopted value of α for the core has little influence on the density distribution within the planet; it is compensated by a change in the value of n since the proportion of helium is adjusted to agree with the empirical mean density. Calculations have been performed on both Jupiter and Saturn with the ratio α for the core put equal to 2, which is 20 per cent smaller than the value (11) previously adopted; the value (9) of α for the mantle has been left unchanged. This change reduces the estimated number n of hydrogen atoms per atom of helium from 13·0 to 10·5 for Jupiter, and from 6·4 to 5·3 for Saturn. The densities in the metallic core are reduced by about 1 per cent in Jupiter and by about 3 per cent in Saturn; the densities in the mantle are increased by about 4 per cent in Jupiter and by about 5 per cent in Saturn. In both planets the radius of the core is altered by less than 1 per cent. The computed moment of inertia of Jupiter is increased by only 1 per cent and that of Saturn by 2 per cent. Thus a change of 20 per cent in the value of the ratio α adopted for the core has little influence on the estimated internal density distributions of the planets. All the uncertain quantities used in the course of these calculations have been examined and no reasonable alterations can appreciably modify the density distributions in Tables V and VI. These density distributions can be altered appreciably only if the assumption of chemical homogeneity is dropped.

The proportion of helium and the central pressure have been determined for each planet from its mass M and its radius R. The only other bulk characteristic of a planet which has so far been determined observationally is the ratio (I/MR^2), where I is the planet's moment of inertia. This ratio is given in Table II for the different planets, and it is 0·25 for Jupiter and 0·22 for Saturn. The computed ratios are somewhat higher. For Jupiter the theoretical density distribution in Table V leads to a ratio (I/MR^2) equal to 0·282, and the ratio computed for Saturn from the data in Table VI is 0·285. These computed values

are fairly close to the empirical values, but they are definitely too large. The discrepancies cannot be attributed to uncertainties in the data used in these calculations; reasonable alterations of the uncertain quantities can change the computed ratios (I/MR^2) by only 1 or 2 per cent. Nor can the discrepancies be ascribed entirely to observational errors; only one-sixth of the discrepancy for Saturn can be accounted for in this way. The computed ratios (I/MR^2) can be lowered appreciably only if the assumption of chemical homogeneity is dropped. Uniform distribution of the heavier elements leads to the maximum possible value of (I/MR^2). The ratio (I/MR^2) is a minimum if the heavy elements are concentrated at the centre of the planet. Mr B. Miles is at present making calculations on Jupiter on the assumption that the heavier elements are concentrated at the centre and he finds that the computed ratio (I/MR^2) is reduced to 0·22. The ratio for Jupiter must therefore lie between 0·22 and 0·28:

$$0{\cdot}22 < (I/MR^2) < 0{\cdot}28. \qquad (13)$$

These comparatively narrow limits are set without any detailed knowledge of the relative abundance of the heavier elements or of their spatial distributions within the planet; it is assumed, however, that the mean atomic weight does not decrease inwards. The ratio 0·25 derived from the empirical data lies halfway between the extreme values (13). It is tempting to conclude that the heavier elements are not uniformly distributed in the planets but are condensed towards the centre. But before a definite conclusion is drawn, the accuracy of the Radau–Darwin formulae (1) and (2) should be tested. These approximate formulae were used to derive the ratio (I/MR^2) from the empirical data. The approximate formulae are no longer necessary since the internal density distributions in the planets can now be estimated. This point will be taken up in a later paper in the hope that the empirical evidence can settle the issue of the chemical homogeneity of the major planets.

Internal temperatures of the planets.—No allowance has been made in the present calculations for the internal temperatures of the planets; the data for hydrogen in Table IV refer to the absolute zero of temperature. Consequently the calculations are valid only if the temperatures in the planets are too low to modify the pressure–density relationship appreciably. The coefficient of thermal expansion of a solid is usually of the order of magnitude of 10^{-5} per deg. C., and so a rise in temperature of 1000 deg. C. decreases the density by only about 1 per cent. The coefficient of thermal expansion is much smaller at high pressures and so correspondingly higher temperatures can be tolerated. The application of a pressure of 10^{11} dynes/cm.^{2}, which is a comparatively small pressure by planetary standards, often decreases the thermal expansion coefficient by an order of magnitude. Moreover, the decrease must continue beyond the experimental range of pressure; this is a direct consequence of the laws of thermodynamics. Bridgman (**26**) has shown from thermodynamics that the absolute magnitude of the derivative $(\partial\rho/\partial T)_p$ at temperature T must decrease at high pressures more quickly than the reciprocal of the pressure p. The coefficient of thermal expansion is therefore unlikely to exceed 10^{-7} per deg. C. in the deep interiors of the major planets. Central temperatures of the order of 10,000 deg. C. would therefore not invalidate the present calculations.

The order of magnitude of the internal temperatures of the planets can be estimated on plausible assumptions. The temperatures in the interior of a planet could be calculated if the flow of heat were known as a function of depth.

In the present calculations it will be assumed that the planets are neither heating up nor cooling down. In other words, it will be assumed that the heat generated by radioactivity in the interior is exactly counterbalanced by the heat conducted to the surface and radiated into space. This assumption may not be rigorously true but it is unlikely to lead to an error in order of magnitude. It has been shown above that Jupiter and Saturn are approximately of solar composition. It is therefore necessary to know the rate at which radioactivity generates heat in solar material. The cosmic abundances of the radioactive elements can at present only be inferred from measurements on meteorites. The average rate of production of heat in meteorites is $1{\cdot}0 \times 10^{-14}$ calorie per gram per second. This is considerably smaller than the average for surface rocks, but the Earth's radioactivity is largely concentrated in the crust (**16**). Indeed, if the Earth as a whole has the same radioactive content as meteorites, only about one-tenth of the heat generated is escaping. But meteorites may have abnormally high radioactive contents since they may originally have formed the surface layers of a planet which "exploded" (**21, 22**). The chemical constituents of meteorites account for only 0·5 per cent of the mass of solar material (**27**) but they include all the radioactive elements. The rate of heating in solar matter will therefore be only 0·5 per cent of that in meteorites, or 5×10^{-17} calorie per gram per second; this figure will be adopted for Jupiter and Saturn. Assuming that heat is radiated from the surface at the same rate as it is generated by radioactivity in the interior, the flow of heat in the outer layers of the planet is

$$1{\cdot}5 \times 10^{-7} \text{ cal. cm.}^{-2} \text{ sec.}^{-1} \qquad (14)$$

for Jupiter and about half this figure for Saturn. The flow of heat in the metallic core of the planet is comparable with that in the non-metallic mantle. But the temperature gradient in the core is very small because of the good thermal conduction. Critchfield (**28**) has made detailed calculations on the thermal conductivity of metallic hydrogen, and he estimates it to be

$$34 \text{ cal. cm.}^{-1} \text{ sec.}^{-1} \text{ deg.}^{-1} \qquad (15)$$

when the density is 1·0 g./cm.^{3}. This thermal conductivity is over thirty times as great as that of the best thermal conductor under laboratory conditions. Moreover, the thermal conductivity of metallic hydrogen increases rapidly with pressure, and it is independent of temperature. It follows from (14) and (15) that the temperature gradient in Jupiter's core is of the order of 4×10^{-9} deg. C. per cm. A temperature difference of 30 deg. C. between the centre of Jupiter and the boundary of the core is therefore sufficient to remove all the heat generated by radioactivity; the temperature difference is even smaller in the core of Saturn. These small differences in temperature are of no importance in the present connection, and the large metallic cores in the planets may be regarded as isothermal. The temperature gradients are much greater in the non-metallic mantles of the planets. The thermal conductivities of rocks and other compact non-metals are usually about 4×10^{-3} cal. cm.$^{-1}$ sec.$^{-1}$ deg.$^{-1}$, and so the temperature differences across the non-metallic mantles of the planets are in the neighbourhood of 30,000 deg. C. The temperature in the metallic cores of the planets are therefore estimated to be about 30,000 deg. C. This estimate of the internal temperatures is likely to be too high for a number of reasons. First, the radioactive content of meteorites may be abnormally large, possibly by an order of magnitude. Second, convection may facilitate the transport of heat across the non-metallic mantles of the planets. Third, the planets may

have been formed cold and are continually heating up. It may therefore be concluded that radioactivity cannot maintain temperatures in the interiors of the planets greater than the order of magnitude of 10,000 deg. C.

Jeffreys (**29**) has examined the thermal histories of the major planets on the assumption that they were originally at stellar temperatures. He has shown that the planets have had ample time to cool by convection. Convection ceases when the temperature gradient falls to the adiabatic gradient which is a few degrees per kilometre. Further cooling by conduction is negligible. Jeffreys therefore concludes that the temperature gradients in the major planets are a few degrees per kilometre and so the central temperatures are of the order of 100,000 deg. C. Jeffreys' argument is valid in the non-metallic mantles of the planets but not in the metallic cores. The thermal conduction in the cores is sufficiently good to remove the heat even in the absence of convection. Jeffreys' estimate of the central temperatures must therefore be reduced by an order of magnitude to 10,000 deg. C. Thus the major planets cannot have central temperatures greater than of the order of 10,000 deg. C., either as a result of radioactivity or as a relic of original heat. The internal temperatures of the planets are therefore too low to modify the pressure–density relationship appreciably.

Discussion.—The principal conclusion of this paper is that the most massive planets in the Solar System, Jupiter and Saturn, are predominantly hydrogen. The hydrogen content of Jupiter is about 80 per cent by mass and that of Saturn about 60 per cent by mass. The hydrogen content of Jupiter has hitherto been put as low as 10 per cent by mass. The proportion of hydrogen in these planets is comparable with that in the Sun and the stars. The hydrogen content in the atmosphere of a star can be estimated spectroscopically and that in the interior can be computed from the bulk properties of the star. Both methods are subject to considerable uncertainties, and recent estimates of the hydrogen contents of the main-sequence stars range from about 40 per cent to over 90 per cent by mass. Nuclear reactions in stars such as the Sun have not appreciably reduced the proportion of hydrogen during geological time. Jupiter and Saturn are therefore not markedly deficient in hydrogen as compared with the original cosmic material. Indeed, the composition of Jupiter would be difficult to reconcile with any but the highest estimates of the hydrogen contents of the stars. It has been shown above that the internal temperatures of the planets are too low to influence the density appreciably and so the pressure–density relationship at absolute zero temperature is applicable. The internal density distributions in Jupiter and Saturn have been estimated on the assumption of chemical homogeneity. The proportion of heavier elements has been chosen for each planet to agree with its empirical mean density. Both planets possess very large metallic cores; their bulk properties are therefore determined mainly by the pressure–density relationship for metallic hydrogen which is known accurately. The computed moments of inertia of the planets are close to, but definitely larger than, the empirical values. This probably means that the heavier elements are not distributed uniformly in the planets but are to some extent centrally condensed; this point requires closer examination.

It is apparent from the above analysis that the planets form three chemically distinct groups. Group I consists of the terrestrial planets, of which the Earth is both the densest and the most massive. The terrestrial planets are of approximately the same chemical composition as meteorites, and their mean atomic weight is over twenty (8). The planets of Group II, Uranus and Neptune, are

more than ten times as massive as the Earth, but they are less than half as dense. It has been shown above that the mean atomic weight in these planets is in the neighbourhood of four. Jupiter and Saturn, which form Group III, are of the order of a hundred times as massive as the Earth and of very low mean density. The hydrogen contents of the planets in Group III are comparable with that in solar material, and their mean atomic weights are little in excess of unity. Thus planets in the different groups differ widely in mass and the more massive planets contain much higher proportions of the lighter elements. The cosmogonical significance of this grouping of the planets is clarified, if not completely explained, by the recent work of ter Haar (**30**), Kuiper (**31**) and Brown (**32**). When the temperature is sufficiently low to permit the formation of molecules, the principal constituents of solar material fall into three well-defined classes. Class I consists of the common terrestrial materials, the most important being metallic iron and the oxides of iron, magnesium and silicon. These materials have high molecular weights and very high boiling-points. The compounds of the second class have molecular weights ranging from 16 to about 20 and boiling-points of the order of 100 deg. K.; the principal constituents in this class are water (H_2O), methane (CH_4) and ammonia (NH_3). The third class consists of hydrogen and helium, which have the lowest molecular weights and also the lowest boiling-points. The compounds of the third class are about 60 times more abundant by mass than those of the second class, which in turn are about four times more abundant than those of the first class (**32**). These three classes include all the important constituents of solar material; in particular, molecular weights intermediate between those of the second and third classes, that is from 5 to 15, are extremely rare. This separation of the constituents of solar material into three well-defined classes explains the division of the planets into three chemically distinct groups. The weak gravitational fields of the terrestrial planets have been unable to retain (or capture) in quantity the light volatile compounds of the second and third classes. The terrestrial planets contain only the highly refractory materials of the first class, and these have probably been retained by chemical forces rather than by gravitation. The observed mean densities of the terrestrial planets are compatible with identical chemical constitutions (**7**); the larger planets are denser because of higher internal pressures. The major planets have been able to retain (or capture) gravitationally large quantities of light volatile materials. Uranus and Neptune appear to have retained only the volatile materials of the second class (water, methane and ammonia); nearly all the hydrogen and helium must have escaped. This is indicated by the mean densities of the planets, which correspond to a mean atomic weight of about four. It also accounts for these planets being an order of magnitude more massive than the Earth; the compounds of the second class are about four times more abundant by mass than terrestrial materials. It is not surprising that Uranus and Neptune have not retained large quantities of free hydrogen and helium. The minimum molecular weight which a planet can retain gravitationally at a given surface temperature is inversely proportional to the gravitational potential (MG/R) at the surface, which is about four times as great for Uranus and Neptune as for the Earth. The Earth has not retained neon, molecular weight 20, which is one of the most abundant elements in the Sun; gaseous elements of even higher molecular weight are markedly deficient on the Earth.* It is therefore understandable that Uranus and Neptune have

* *Cf.* Harrison Brown, reference (31), chapter ix.

not retained in quantity compounds whose molecular weights are less than 5. On the other hand, the gravitational potential (MG/R) at the surface of Jupiter is about 30 times as great as on the Earth, and on Saturn it is about ten times as great. It is therefore not surprising that Jupiter and Saturn have retained (or captured) considerable proportions of the original hydrogen and helium, as the calculations of the present paper indicate. The retention (or capture) of the light materials of the third class explains both the large masses of Jupiter and Saturn and their very low mean densities. Cosmogonical considerations along these lines have been examined in greater detail by other workers (**30, 31, 32**), and their main conclusions are compatible with the results of the present investigation. To summarize, it may be stated that the bulk properties of a planet are determined primarily by its mass. This is the simplest explanation of the differences between the planets, and it is consistent with all empirical data. The chemical composition of a planet has been determined by its ability to retain or capture light volatile materials and the mass of the planet is the governing factor. The planets are divided by mass into three distinct groups, according to their abilities to retain (or capture) the three principal classes of compounds in solar material. Planets in the same groups are of approximately the same chemical composition, and the more massive planets are denser because of greater compression.

Uranus and Neptune are presumably composed mainly of water, methane and ammonia, together with terrestrial materials. A quantitative treatment of the bulk properties of these planets must therefore await the results of calculations on the pressure–density relationships of these compounds. A quantitative treatment would also require more accurate empirical data. In particular, the diameter of Uranus should be re-determined by Kuiper's (**14**) new technique. Reliable estimates of the radii of these planets are of especial interest since it seems likely that the more massive planet, Neptune, has the smaller radius (*cf.* Table I). This situation would be very difficult to reconcile with the earlier models of the planets and it seems to be associated with a transition to a metallic phase. Normally, the radius of a planet of given chemical composition is an increasing function of its mass. The more massive planets may contain higher proportions of the lighter elements and this would exaggerate the increase of radius with mass. The radius could, however, suddenly decrease with increase of mass if the density increases abruptly by more than 50 per cent at a certain critical pressure; this phenomenon has been discussed at length in other papers (**21, 22**). The radius of Neptune could be smaller than that of Uranus if one of the principal constituents (water, methane or ammonia) became metallic at some pressure intermediate between the central pressures of the two planets. This would be a striking demonstration of the importance of transitions to metallic phases at planetary pressures. The critical pressures necessary to make water and methane metallic are likely to be high, and one of them might lie between the central pressures of Uranus and Neptune. On the other hand, common ammonia (NH_3) may not be important in the deep interiors of the planets; the nitrogen may be combined mainly in the form of metallic ammonium (NH_4^+). Ammonium metal has never been prepared in the laboratory but its existence under suitable conditions has long been taken for granted. The ammonium ion (NH_4^+) is a stable chemical unit which has many similarities to the ions of the alkali metals. Also, ammonium forms an amalgam which is really an alloy

of metallic ammonium and mercury. The metal NH_4 would be unstable under laboratory conditions; it would dissociate into ammonia (NH_3) and hydrogen (H_2). The tendency to dissociate would disappear at sufficiently high pressures; the metal occupies a smaller volume than the dissociation products. The critical pressure above which the metal NH_4 is stable is probably low by planetary standards, possibly of the order of magnitude of 10^{11} dynes/cm.^{2}. An attempt is at present being made at University College, London, to calculate the pressure–density relationship of metallic ammonium, and these calculations may be a first step towards a quantitative treatment of the bulk properties of Uranus and Neptune.

A large part of this investigation was carried out at the Institute of Theoretical Physics in Copenhagen, and the author would like to thank Professor N. Bohr for his hospitality. He would especially like to express his appreciation to Professor L. Rosenfeld and to Professor Harold Jeffreys for the interest which they have shown in this series of investigations and for their many helpful suggestions.

The Physical Laboratories,
The University,
Manchester, 13:
1951 *May* 9.

References

(1) Rupert Wildt, *M.N.*, **107**, 84, 1947.
(2) Harold Jeffreys, *M.N., Geophys. Suppl.*, **4**, 50, 1937; **4**, 62, 1937.
(3) J. D. Bernal, *The Observatory*, **59**, 268, 1936.
(4) W. Kuhn and A. Rittmann, *Geol. Rundsch.*, **32**, 215, 1941.
(5) R. Kronig, J. de Boer and J. Korringa, *Physica*, **12**, 245, 1946.
(6) Harold Jeffreys, *Nature*, **165**, 218, 1950.
(7) W. H. Ramsey, *M.N.*, **108**, 406, 1948.
(8) W. H. Ramsey, *M.N., Geophys. Suppl.*, **5**, 409, 1949; **6**, 42, 1950.
(9) K. E. Bullen, *An Introduction to the Theory of Seismology*, Cambridge, 1947.
(10) Harold Jeffreys, *M.N.*, **94**, 823, 1934; *M.N., Geophys. Suppl.*, **4**, 62, 1937.
(11) Rupert Wildt, *Ap. J.*, **87**, 508, 1938.
(12) D. S. Kothari, *Proc. Roy. Soc.* A, **165**, 486, 1938.
(13) H. N. Russell, R. S. Dugan and J. Q. Stewart, *Astronomy*, Boston, 1945.
(14) G. P. Kuiper, *Ap. J.*, **110**, 93, 1949.
(15) N. Lvoff, *Russ. Astr. J.*, **9**, 68, 1932.
(16) Harold Jeffreys, *The Earth*, Cambridge, 1929.
(17) Harold Jeffreys, *M.N.*, **84**, 534, 1924.
(18) B. M. Peek, *M.N.*, **97**, 574, 1937.
(19) W. H. Ramsey, *M.N.*, **110**, 444, 1950.
(20) F. Seitz, *The Modern Theory of Solids*, New York, 1939.
(21) W. H. Ramsey, *M.N.*, **110**, 325, 1950.
(22) M. J. Lighthill, *M.N.*, **110**, 339, 1950.
(23) H. and B. S. Jeffreys, *Methods of Mathematical Physics*, Cambridge, 1946.
(24) J. E. Lennard-Jones, *Physica*, **4**, 941, 1937.
(25) J. G. Scholte, *M.N.*, **107**, 237, 1947.
(26) P. W. Bridgman, *The Physics of High Pressure*, London, 1949.
(27) R. v. d. R. Woolley and C. W. Allen, *M.N.*, **108**, 292, 1948.
(28) C. L. Critchfield, *Ap. J.*, **96**, 1, 1942.
(29) Harold Jeffreys, *M.N.*, **83**, 350, 1923; **98**, 214, 1938.
(30) D. ter Haar, *Det Kgl. dansk. Vid. Selsk., Mat.-fys. Medd.*, **25**, No. 3, 1948.
(31) G. P. Kuiper, *The Atmospheres of the Earth and the Planets*, Chicago, 1949, chapter xii.
(32) Harrison Brown, *Ap. J.*, **111**, 641, 1950.

THE COMPUTATION OF TOPOCENTRIC LIBRATIONS

R. d'E. Atkinson

(Received 1950 October 13)

Summary

A short method is given for computing components of differential libration, so that the geocentric values tabulated in the Ephemeris can be converted to those for an observer anywhere, instead of computing the total values *ab initio*. The method makes slide-rule accuracy amply adequate, where four-figure tables were slightly inadequate before, and it also comprises appreciably fewer steps. The differential correction to the apparent position-angle of the Moon's axis is shown to be necessary, and a formula for it is given. The question of the accuracy actually needed is considered, and it is shown that changes of parallax affect the limb-contour as much as do changes of libration of several units in the last place tabulated at present, so that one might almost speak of a " parallactic libration " in addition to the other two. All three must be specified if the present tabular accuracy in the two standard ones is to be justified.

It is well known that a severe limit is set to the accuracy with which occultation and eclipse observations can be effectively reduced, so long as the lunar limb-detail has to be ignored. The limb-contour is strongly dependent on the librations in both longitude and latitude, and indeed the values for each point on the apparent limb depend on both of these librations separately, and thus form a two-dimensional array, so that a large mass of data must be assembled before the limb-regions can be well enough known for practical use. Up to the present, corrections have generally not been applied, simply for lack of this information, and the question whether the computation of the librations could itself be simplified has scarcely arisen.

This situation is now changing; elaborate and systematic knowledge of the contours will soon become generally available, as a result of the extensive programme being undertaken by C. B. Watts at the Naval Observatory in Washington, and a very large number of librations will in future have to be computed, for the circumstances of individual observers of occultations; the question whether any saving is possible in the computational method to be used has thus become important even though the maximum attainable saving in any one case may be fairly small.

The librations are published in the *Nautical Almanac* only for a geocentric observer, which means in effect an observer for whom the Moon is in the zenith. They range, roughly speaking, over about $\pm 7\frac{1}{2}^\circ$ in (selenographic) longitude, and slightly less in latitude. For any other observer, the values will differ from the geocentric ones according to the parallactic corrections which he must apply to the Moon's tabular apparent place; these differential corrections may themselves approach one degree, and it is quite evident that they can neither be guessed to within a very few hundredths of a degree, nor be ignored if that standard of accuracy is desirable, as in fact it is: computation is essential. The procedure

outlined in the *Nautical Almanac* is not to calculate these corrections as such at all, but to start from scratch and recalculate the entire topocentric librations just as though the geocentric values were not available. The method, described in the " Explanation " for 1940, is to compute the total parallax and the parallactic angle, for the observer's position at the time; resolve the parallax into its components in R.A. and Dec., and apply these to the Moon's tabular place; convert the apparent coordinates so obtained into apparent longitude (λ) and latitude (β), by the usual formulae; obtain ☾, the Moon's mean longitude, from the tables, and derive $\lambda - ☾$; apply corrections obtained from a " Table for the Optical Libration of the Moon " which has to be entered with the argument $\lambda - ☊$; and finally add the physical librations, which are tabulated separately. Since the leading term in the longitude-libration is the quantity $\lambda - ☾$, this subtraction will in itself normally destroy one significant figure; owing to this, and to the considerable number of other steps, four-figure tables will not regularly be sufficient for a final accuracy of 0°·03, even though the answer never exceeds 10°. The accuracy of the tabulated geocentric values is 0°·01, and the former figure (as will be shown below) is not really adequate for the best observations now obtainable; the method thus demands five-figure tables.

The purpose of the present paper is to develop the alternative approach, and give formulae for differential corrections to be applied to the geocentric values already tabulated in the *Almanac*. This will naturally reduce the number of significant figures which must be carried, but in fact it also materially reduces the actual number of steps in the operation. As far as the number of significant figures is concerned, the entire amplitude of the differential libration, before resolution into longitude and latitude components, does not exceed 0°·995, even at perigee, if the Moon's altitude is greater than 16°, and throughout most of the month it is below this figure at all altitudes, so that we can say that the differential method practically never demands more than two significant figures in the answer; the number of steps is so small that rounding-off errors will very seldom exceed 0°·02 if only two figures are worked throughout, and slide-rule accuracy is therefore amply adequate in all cases. The method of procedure is as follows.

The libration is normally specified by stating the selenographic longitude and latitude of the apparent disk-centre at the time in question, measured from the mean centre as origin. It could equally well be characterized by the distance from this mean centre to the apparent one (either in kilometres on the Moon's surface or in seconds at the Earth's centre), and a suitable position-angle. For a non-geocentric observer, the location of the apparent disk-centre will differ from that for a geocentric one, both in magnitude and direction, simply by the amount of his own parallax; it is thus only necessary to compute this parallax, and to resolve it into components parallel to the directions of selenographic latitude and longitude at the centre of the Moon's apparent disk, in order to obtain differential corrections to add to the values in the Ephemeris. Even when the libration in latitude is not zero, it is clear that this means resolving parallel and perpendicular to the projection of the Moon's axis on the celestial sphere, as judged by the geocentric observer. The longitude-component obtained in this way must of course be multiplied by $\sec b$, where b is the total libration in latitude, if the result is to be in degrees of lunar longitude; $\sec b$, however, is never as large as 1·008, and can be allowed for mentally. (In eclipses, it is always 1·0000 at the region of totality.)

In Fig. 1, E is the centre of the Earth, O is an observer on the Earth's surface, and M is the Moon's centre. π is the Moon's equatorial horizontal parallax (tabulated in the *Almanac* at half-day intervals), and π' is the parallax for the observer, i.e. the total amount by which his view of the Moon differs from the geocentric view. Neglecting the ellipticity of the Earth, the angle OEM, or ζ, is evidently the zenith distance, at O, of the Moon's geocentric place, i. e. the Moon's zenith distance uncorrected for parallax; and taking the Earth's radius as unity the distance EM is $\operatorname{cosec}\pi$. We thus have

$$\tan\pi' = \frac{\sin\zeta}{\operatorname{cosec}\pi - \cos\zeta}$$
$$\approx \sin\zeta\sin\pi(1+\cos\zeta\sin\pi) \qquad (1)$$

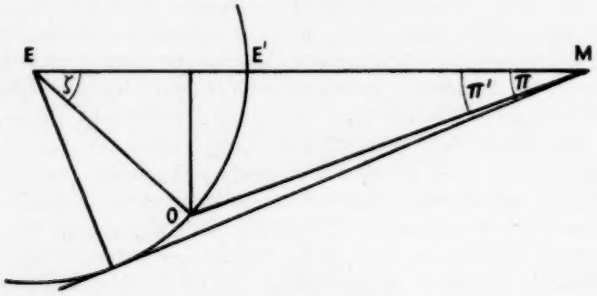


FIG. 1.

very nearly, since $\cos^2\zeta\sin^2\pi$ cannot reach 0·0004. In fact, since π and π' are both small, while $\sin\pi$ varies only between about 0·0157 and 0·0180, it is sufficient to write

$$\pi' = \pi\sin\zeta(1+0{\cdot}0168\cos\zeta)$$
$$= \pi(\sin\zeta + 0{\cdot}0084\sin 2\zeta), \qquad (2)$$

where π may be in any convenient units, and π' will be in the same. The quantity ζ may be obtained, in the usual way, from

$$\cos\zeta = \sin\delta\sin\phi + \cos\delta\cos\phi\cos H, \qquad (3)$$

where H and δ are the Moon's tabular hour angle and declination and ϕ is the latitude, all for the point O, and without any allowance for parallax. For any given observatory, of course, tables giving the zenith distance in terms of H and δ may already exist; general altitude-azimuth tables are also serviceable. It will normally be sufficient to obtain ζ to the nearest quarter degree, even if it is small.

To resolve π' into its components in the directions of selenographic longitude and latitude, we proceed as follows. Fig. 2 shows the Moon's disk as viewed by the observer for whom it is in the zenith (i. e. looking along $EE'M$ of Fig. 1). MN is directed to the celestial pole; the figure is drawn for the case that O and M are in the northern hemisphere, with the Moon's hour-angle at O positive, and with C_0, the position-angle of the Moon's axis (RT), also positive. (The north pole of the Moon will be at R if the libration in latitude is zero; in other cases, it will be very slightly inside R, but projected on the line RT.) $MO'O$ is the great circle passing through the zenith of O, and the shift of the observer from E' to O (Fig. 1), through the angle ζ at the Earth's centre, shifts the apparent

disk-centre by the small distance $\pi' r$ on the Moon's surface, r being the Moon's radius. The displacement is from M towards O (Fig. 2), and it will never exceed $MO'/57{\cdot}3$; its components in selenographic longitude and latitude are evidently

$$\left.\begin{aligned} \Delta l &\approx -\pi' \sin(Q - C_0) \sec b \\ \text{and} \quad \Delta b &\approx +\pi' \cos(Q - C_0) \end{aligned}\right\} \tag{4}$$

respectively. The angle Q, which is the azimuth of O at the point where the Moon is in the zenith, is of course also the parallactic angle for the Moon (uncorrected for parallax) at O, and may be obtained from

$$\sin Q = \sin H \cos\phi \operatorname{cosec}\zeta \tag{5}$$

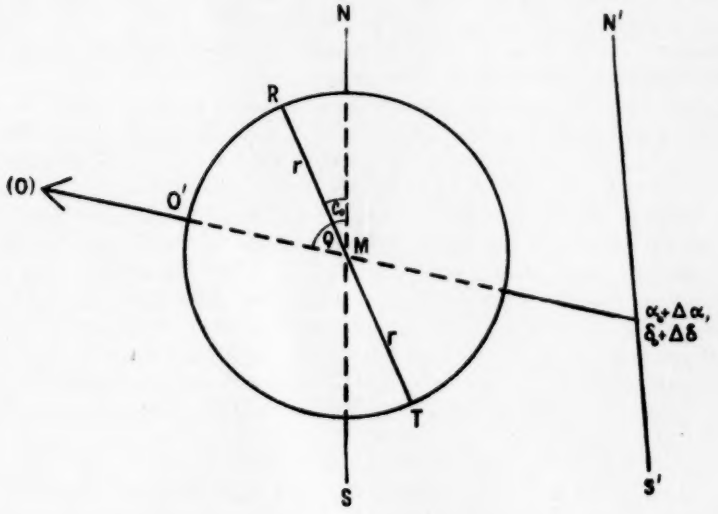


FIG. 2.

if Q is not too close to 90°; alternatively, it may be obtained from

$$\cos Q = \frac{\sin\phi - \cos\zeta \sin\delta}{\sin\zeta \cos\delta} \tag{6}$$

or from tables as before. The total librations themselves, l and b, are then obtained from

$$\left.\begin{aligned} l &= l_0 + \Delta l, \\ b &= b_0 + \Delta b, \end{aligned}\right\} \tag{7}$$

where l_0 and b_0 are the geocentric librations, including the physical libration, tabulated in the *Almanac* for every day. Second differences will be needed in interpolating these two quantities; but not for C_0 (at this stage), and still less for any of the other quantities required.

The procedure may be recapitulated, in the working order, as follows:—

A. Obtain from the Ephemeris the geocentric R.A. and Dec. for the Moon, say to the nearest half minute of time and tenth of a degree of arc. Obtain the local sidereal time to the same accuracy, the Moon's parallax to the nearest half

minute, and the position-angle of the Moon's axis to the nearest fifth of a degree. Obtain also the latitude of the observer, to the nearest tenth of a degree.

B. Equations (3), (5) (or (6)), (2), and (4) are then taken in this order; tables may replace (3), (5), and (6) if preferred.

C. Obtain l_0 and b_0 as accurately as possible by interpolation in the Ephemeris ("Earth's selenographic longitude and latitude"), and use (7).

The only steps in the above work which are not required in the standard method also are the interpolation of C_0, l_0 and b_0; (equations (4) are replaced by two very similar ones). All the other steps are not only required but in several cases are required to appreciably greater accuracy, and the other steps described above in the summary of that method must be added also. It is clear that the total amount of labour is appreciably greater than with the differential method, and that it would be advantageous to include the above precepts A, B and C in the *Almanac*; in that case, it would apparently also be permissible to drop the "Table for the Optical Libration" entirely.

The question may be raised whether the accuracy of $0^\circ{\cdot}01$ in l and b, contemplated above and provided in the Ephemeris for l_0 and b_0, is necessary, or sufficient. If it is not, the limits of accuracy in the working will have to be modified for both methods, but the differential method will still retain about the same lead over the other. So far as limb-corrections are concerned, it would be desirable in principle to provide for apparent lunar height-differences down to $0''{\cdot}01$ at any rate for use in photoelectric occultation work; down to say $0''{\cdot}03$ for eclipse work; and probably down to $0''{\cdot}05$ or less for direct photographic determinations of the limb-contour, such as are now being carried out systematically in Washington. Even if we concede that eclipses and occultations will not of themselves provide limb-contours, the value of $0''{\cdot}05$ remains valid for them, for the third reason; it may, indeed, already be rather insufficient. The radius of the Moon being 1738 km., and mountains on and near the limb up to $2''$, or 3·7 km., above the mean level being fairly common, it is clear that a mountain may contribute to the limb-detail even if it is as much as 114 km. this side of, or beyond, the ideal "limb"; if the Moon's surface near the ideal limb happens to have a depression at the position-angle in question, the mountain can be seen appreciably farther still, and as there are also peaks exceeding $2''$ in height we may at least extend the limit to 130 km. Since $0''{\cdot}05$ corresponds to a height-change of 86 metres at the Moon's least distance, the libration must be known at any rate to

$$\frac{86 \times 57^\circ{\cdot}3}{130000} < 0^\circ{\cdot}04;$$

however, in order to allow of interpolating, so as to obtain a contour for a given occultation (say) from four known contours whose centres surround the disk-centre required, the individual values should preferably be known to finer limits than this. It should be stressed that, especially when two mountains are nearly in line, one on the near side of the limb and one beyond it, the contour of the east and west limbs is quite sensitive to the libration in latitude, as well as to that in longitude; and similarly the north and south limbs are sensitive to that in longitude, as well as latitude. The accuracy of $0^\circ{\cdot}01$ is thus suitable at present, though perhaps not yet fully required; an error of $0^\circ{\cdot}03$ could introduce changes comparable with the accuracy of detail in the Washington work, and definitely larger than any purely experimental error in photoelectric occultation work, so

that the total computational error should be kept definitely below this figure, and preferably down to 0°·01.

We turn now to the question of the position-angle, C, of the Moon's axis, i.e. of that lunar meridian which bisects the apparent disk and so appears as a straight line. This, also, is tabulated in the *Almanac* for a geocentric observer only, and the question arises whether differential corrections are needed in occultation work. (Other lunar observations, such as heliometer ones for improvement of the Moon's equatorial inclination, or of the physical librations, may be treated by direct reductions in full; such cases are relatively rare.) Some traces of the limb have been very kindly supplied to the writer by Mr C. B. Watts for another purpose, and these contain a small but not negligible number of regions in which the contour rises at slopes as steep as about $\tan^{-1}0{\cdot}2$ for distances longer than 86/0·2 metres. Thus if the limit of accuracy of 86 metres in height, as used above, is not to be exceeded on account of errors in position-angle, these errors should not exceed $dC = \frac{86}{0{\cdot}2} \times \frac{57{\cdot}3}{r}$ degrees, where r is the Moon's radius in metres. This gives $dC = 0{\cdot}014$ degrees, and it is easily verified that the geocentric values are not applicable, with this accuracy, for arbitrary positions of the Moon and the observer. Since however, such steep regions are rare, a differential correction which is itself reliable to 0·02 degrees will in general be adequate.

The differential correction to the position-angle may be separated into two parts. On the one hand, there is the contribution due to convergence of the meridians on the celestial sphere; as the observer's movement alters the Moon's apparent place from (α_0, δ_0) to $(\alpha_0+\Delta\alpha, \delta_0+\Delta\delta)$ (see Fig. 2), the line joining the north extremity to the south extremity, whose position-angle was of course zero to begin with, takes up a position-angle (referred to the new meridian $N'S'$), $\Delta_1 C = \Delta\alpha \sin\bar{\delta} = -\pi' \sin Q \tan\bar{\delta}$, where $\bar{\delta} = \delta_0 + \frac{1}{2}\Delta\delta$, even though the line has only been displaced parallel to itself. In addition, if b, the libration in latitude, is not zero, the displacement Δl will bring this forward tilt into view from the side, to a small extent; the change in position-angle due to this effect is evidently $\Delta_2 C = \sin b \,.\, \Delta l$. To the order of accuracy required, these are the only two components that need be considered. We thus have

$$\Delta C = \sin b\, \Delta l - \pi' \sin Q \tan\bar{\delta} \qquad (8)$$

where b is to be taken literally as $b_0+\Delta b$ (which may evidently be done without any trouble). $\bar{\delta}$ is better than δ_0, but the difference, $-\frac{1}{2}\pi' \cos Q$, cannot exceed half a degree and may be omitted without appreciable effect; it may also readily be estimated to the nearest fifth of a degree, which is amply good enough. The corrected position-angle C is then obtained from

$$C = C_0 + \Delta C, \qquad (9)$$

where C_0 is the tabulated geocentric value, interpolated now as accurately as possible, with second differences.

If we work to the accuracy of 0°·03 there is one further aspect of libration to be considered in limb-contour work (including observations of Mösting A against the limb). If the parallax changes, then even though l and b do not change, the appearance of the limb may; if the parallax increases, mountains beyond the ideal limb will be lowered, whatever their position-angles, and

mountains nearer than the ideal limb will be raised. This might almost be described as the "parallactic libration".

Fig. 3 shows two positions, M_1 and M_2, for the Moon, the nominal librations for the observer at O being supposed the same in both cases; it is clear that the selenographic coordinate, θ, of the ideal limb (the point L where the tangent touches the circle) and the direction of the tangent itself have both changed by the amount $s_1 - s_2$, where s is the apparent semi-diameter, just in the same way

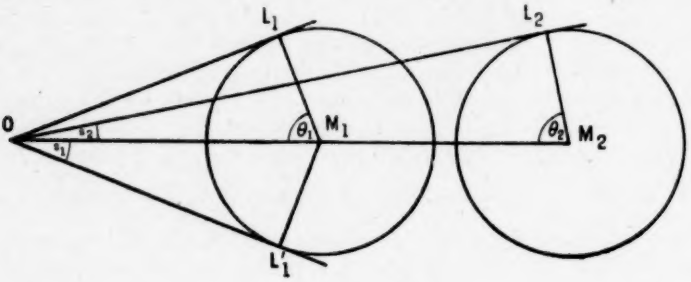


FIG. 3.

(for this particular point of the limb) as if the Moon had actually suffered a change of libration of $s_1 - s_2$ in the plane of the figure. (For the point L', the sign of this "libration" is of course opposite to that for L, and its resolution into longitude and latitude varies continuously with position-angle.) Since $s_1 - s_2$ may in fact be as large as $0^\circ{\cdot}039$ approximately, the effect cannot be entirely neglected, even at present.

It remains only to verify that the ellipticity of the Earth can safely be neglected, as has been done in the above discussion. The ellipticity displaces the point of observation, from what it would be if the Earth were spherical and ρ and ϕ were used unchanged, by absolute space-distances whose maximum value is not more than about 30 km.; at the Moon's least distance, 30 km. subtends an angle of

$$\frac{30 \times 57^\circ{\cdot}3}{356000} = 0^\circ{\cdot}0048.$$

If this is to be included at all, it can be adequately done by very rough critical tables with argument ϕ.

I wish particularly to express my thanks to Mr C. B. Watts, of the U.S. Naval Observatory, for several most helpful letters.

Royal Observatory,
Greenwich:
1950 *October* 12.

EXPANDING WORLD-MODELS CHARACTERIZED BY A DIMENSIONLESS INVARIANT

G. J. Whitrow and D. G. Randall

(Received 1951 May 21)

Summary

A class of homogeneous expanding world-models, which includes Milne's uniformly expanding model as a particular case, is analysed by the methods of kinematic relativity, red-shifts being attributed in each case to the Doppler effect associated with mutual recession. It is found that each model (of the class) is characterized by the invariance of the dimensionless product $G\rho t^2$, where G is the constant of gravitation, ρ the mean local density and t the parameter occurring in " Hubble's law " (which holds good locally for each model). The " present " ages and (local) densities of several of these models are determined for comparison with the results of observation of the extra-galactic nebulae.

1. *Introduction.*—In 1946 one of us (**1**) showed that a simple argument based on Newtonian gravitational theory and Einstein's relation between mass and energy leads to a result which is almost identical with the well-known formula correlating the mass M and radius R of the Einstein universe. In 1949 (and previously) P. Jordan (**2**) developed independently a cosmological theory in which one of the main ideas leads by a similar argument to an equivalent result. If in Einstein's relation we formally substitute ct for R, we obtain an equation which is very nearly the same as that obtained by Milne (**3**) for the substratum in his cosmology, viz.

$$GM = c^3t, \tag{1}$$

where G is the classical constant of gravitation, M the so-called "fictitious mass" of the substratum (in Milne's theory, as presented by him, the substratum is infinite), being the mass of a Euclidean sphere of the same uniform density as the local density of the substratum and of radius ct. In this world-model t is the "age" of the system, i.e. the local cosmic epoch, and c is the velocity of light. This remarkable, and unique, point of agreement between world-models otherwise so widely different (in Einstein's theory there is no expansion and no creation of matter, in Milne's there is expansion but no creation *, whereas in Jordan's there is both expansion and creation) was originally ascribed by the author of (**1**) to Mach's principle, interpreted as the identification of inertia and world-gravitation; whereas Jordan (**2**) has explicitly adopted the effectively equivalent postulate that the total energy of the universe, i.e. the algebraic sum of the rest-energies of all material particles and their (negative) gravitational potential energies, is exactly zero.

Equivalent formulae to those of Einstein, Milne, Jordan and Whitrow, involving fewer parameters, arise if we replace M by expressions involving the local density ρ. Thus, Milne's relation (**1**), for example, gives

$$G\rho t^2 = (4\pi/3)^{-1}. \tag{2}$$

* Since, strictly speaking, only epochs *subsequent* to $t=0$ can be considered in this theory.

In comparing his theory with observation, ρ is "identified" with the local smoothed-out density of the extragalactic nebulae derived from nebular counts, and t is interpreted as the empirical parameter of time t_0 which arises when Hubble's red-shift law for the nearer nebulae is expressed in the form *

$$\delta\lambda/\lambda = r/ct_0, \tag{3}$$

where r denotes the estimated distance of a nebula. Since G is approximately $6{\cdot}67 \times 10^{-8}$ in c.g.s. units and Hubble's value for t_0 is approximately 6×10^{16} sec., the corresponding theoretical value of ρ, as determined by (2), is of the order of 10^{-27} g./cm.^{3}. This value is generally regarded as being *at least* ten times as great as the highest value which can be derived from observation. With regard to this discrepancy, the following comments may be made:—

(i) It is significant that the observational value now generally accepted (10^{-28} to 10^{-29} g./cm.^{3}) is about a hundred times larger than that put forward twenty years ago and *may* also, in its turn, prove to be an underestimate.

(ii) The local group comprises three large bright nebulae and about a dozen much smaller and less luminous nebulae, so that it is probable that outside this group there exist many more small faint nebulae than previously suggested.

(iii) The problem of estimating the masses of extragalactic nebulae is notoriously difficult, and the principal methods hitherto employed—namely, by applying the virial theorem to clusters of nebulae, by considering spectrographic rotations, and by considering velocities of pairs of nebulae—give masses differing by at least a factor of ten for nebulae of the same absolute magnitude.

(iv) In discussing the difficulties of (iii) at the I.A.U. meeting in 1948 at Zürich (**4**), it was agreed that "intergalactic matter within the great clusters of nebulae cannot definitely be ruled out as a contributing factor to explain part of this discrepancy", and it is possible that far more intergalactic matter exists than has previously been assumed.

(v) The recent discovery of radio stars in our own and neighbouring galaxies, if confirmed and extended, *may* ultimately provide further evidence for the belief that in the universe as a whole considerably more matter exists (in an effectively non-luminous state) than was formerly supposed.

A further difficulty which arises when Milne's substratum is compared with observation concerns his identification of the parameter t_0 given by (3) with the present value of the parameter t occurring in (2). The latter parameter is strictly the radius of the model divided by c, and as it is assumed that the model is expanding uniformly t is also equal to the time. Thus, in this system it is assumed that Hubble's parameter t_0 changes with epoch and in effect measures the epoch, so that the present age of the system is approximately 2×10^9 years. This estimate for the "age of the universe" is now thought to be on the low side, the fashionable estimate at present being about 5×10^9 years, i. e. not less than two and possibly three times the epoch given by Hubble's parameter t_0 occurring in (3).

Despite these possible discrepancies, formula (2) is of considerable intrinsic interest. First, we note that if we associate with this formula a lower value of ρ than that adopted (in theory) by Milne we must assign at the same time a greater value to t, e. g. if ρ is to be diminished by a factor ten then t must be increased by a factor slightly greater than three. Furthermore, the formula, whether strictly

* See E. P. Hubble, *Observational Approach to Cosmology*, Oxford, p. 37, 1937, where this law appears in the form $\delta\lambda/\lambda = kr$. Our parameter t_0 is *defined* by the equation $k = 1/ct_0$.

applicable to the physical universe or not, focuses attention on two important points:—

(i) The product of three empirically determined quantities, namely, the classical constant of gravitation, the local smoothed-out density of the system of extragalactic nebulae and the square of Hubble's parameter t_0, as defined by (3), is of zero physical dimension, i. e. is a pure number.

(ii) The empirical value of this number is not *widely* different from unity, e. g. it lies between about 2×10^{-1} and 2×10^{-3}, according as the local density is taken to be any value between 10^{-27} and 10^{-29} g./cm.^{3}.

These results suggest that a formula of the type (2) may be of wider significance than the particular model of Milne in which it first arose. In this model G, ρ and t_0 all vary with lapse of time, but the product of the first two and the square of the third is a dimensionless invariant, which must express some permanent feature of the model, and it is clear from the expression occurring on the right-hand side of (2) that this feature is simply its three-dimensional spatial character.

More generally we may look for world-models admitting a relation similar to (2), the coefficient of $(4\pi/3)^{-1}$ on the right-hand side not necessarily being exactly unity. In such models it will not happen as a rule that the value of "Hubble's parameter" t_0 at any given epoch t is exactly equal to t, as occurs in the uniformly expanding model of Milne, although there may be some simple invariant relation between them. Although, in the models to be considered, "Hubble's parameter" will change with epoch, it will be necessary to denote it by a different variable from that denoting the epoch. Henceforward, we shall use the symbol t exclusively to denote a parameter occurring in the model under consideration which plays a rôle in that model analogous to the empirical rôle of Hubble's parameter, i. e. it will be given by r/cs, where r denotes the distance and s the corresponding theoretical red-shift ratio (fractional change of wave-length) of a *neighbouring* nebula in the model. In comparing the model with observations, we will take Hubble's empirical value of t_0 to be the "present" value of this parameter. Further, we shall use the symbol T to denote the epoch, which in general will be equivalent to the age of the model. When there is a simple relation of constant proportionality between t and T then the invariance of either of $G\rho t^2$ or $G\rho T^2$ will involve that of the other.

In classical Newtonian theory and in the standard world-models based on General Relativity, G is a true constant of nature which does not vary with time. In general, for the models we shall consider, the parameter playing a rôle analogous to the constant of gravitation will be found to change with epoch, being proportional to some characteristic power of the epoch, but the value to be assigned to it at the present time will of course be the standard empirical value $6{\cdot}67 \times 10^{-8}$ in c.g.s. units.

The object of the present paper will be the construction of a system of world-models, including Milne's substratum as a particular case, in each of which the constituent fundamental particles will be all equivalent and the model as a whole will be homogeneous and isotropic. The metrology of each model will be based on Milne's convention concerning the measurement of distance in terms of time readings and an invariant velocity of light (signal velocity) c. Each model will be an expanding one, the theoretical red-shifts arising being attributed solely to the effect of mutual recession and so being essentially Doppler shifts. We shall examine how G, ρ and t in these models depend on the epoch T, and we shall find

that in all cases considered relations analogous to (2) are satisfied. In comparing the models with observation we will calculate the current predicted values of ρ and T corresponding to the current observed values of G and t, and see what conclusions can be drawn concerning present estimates of the mean density and age of the universe if it is assumed to be continually expanding.

2. *A class of world-models.*—It is well known (**5**) that all members of a collinear set of observers are equivalent in pairs if the signal functions $\theta_{pq}(T)$ correlating the epochs of reception and emission of signals passing between them can be expressed in the form

$$\theta_{pq}(T) = \Omega\{\Omega^{-1}(T) + \lambda_{pq}\}, \tag{4}$$

where p, q are suffixes denoting a particular pair of observers, λ_{pq} is a constant for given p, q, and Ω is a functional operator characterizing the whole set of observers and so independent of p, q. Beginning with a particular observer O, we can construct a pencil of sets of observers of this type, each ray of the pencil corresponding to each spatial direction at O. When $\Omega(T) \equiv T$, a signal always takes the same "time" λ_{pq} to pass from one given observer to another, and with appropriate conventions of measurement (**6**) O will describe all the other observers as relatively stationary. Each observer will describe the general motion-pattern of the whole system in the same way if the geometry of the whole system is of constant (or zero) curvature, the signal paths connecting the observers corresponding to the geodesics. If we regard each observer as an origin of measurement, signalling, etc., associated with a "particle", then the corresponding system of particles can define a homogeneous world-model. Without further discussion, we shall assume from now on that our systems are of this type.

So far we have defined a model only for the case $\Omega(T) = T$, but the general case given by (4) can always be transformed mathematically into this special case by regraduating all clocks from T to τ, where

$$\tau = \Omega^{-1}(T), \qquad T = \Omega(\tau). \tag{5}$$

If the description of the whole system is uniform and isotropic according to each observer after this mathematical transformation, then it must be so beforehand, since the measurements, etc., of each observer are changed according to the same rules. Thus it is sufficient to postulate that the world-model under discussion reduces to the type defined in the previous paragraph when submitted to the transformation (5). It may be mentioned that, so far as the present paper is concerned, we regard (5) purely as a legitimate *mathematical* process, and we need make no appeal to any physical postulate concerning the coexistence of different scales of time.

In order to specify a particular class of models of this type we shall now place some restriction on the function Ω. This function is usually assumed to be single-valued and monotonic increasing, so that it has a unique inverse which is also monotonic increasing. It is useful to associate with it a function $\omega(T)$ defined by the relation

$$\Omega^{-1}(T) = \int^{T} \frac{dT}{\omega(T)}, \tag{6}$$

the lower limit of integration being any suitable arbitrary constant. The function $\omega(T)$ serves to define the infinitesimal operator of the continuous one-parameter group of functional operators θ_{pq} corresponding to a given Ω. To each $\omega(T)$ there

will correspond, in general, a particular type of model. We shall consider the cases in which $\omega(T)$ is proportional to a positive (or zero) power of T, i.e.

$$\omega(T) = kT^n, \tag{7}$$

where $n \geqslant 0$ and $k > 0$. Equation (6) then gives, if $n \neq 1$,

$$\Omega^{-1}(T) = \frac{1}{k(n-1)}\left(\frac{1}{A^{n-1}} - \frac{1}{T^{n-1}}\right), \tag{8}$$

where A is a constant of integration, and hence

$$\Omega(T) = \{A^{1-n} + k(1-n)T\}^{1/(1-n)}. \tag{9}$$

Consequently, writing λ for λ_{pq}, equation (4) gives

$$\theta(T) = \{T^{1-n} + k(1-n)\lambda\}^{1/(1-n)}. \tag{10}$$

For dimensional consistency, we take λ to be of the dimensions of time and k to be of the dimensions of time raised to the power $-n$. Formulae (8), (9) and (10) do not apply in the singular case $n = 1$. We then have

$$\Omega^{-1}(T) = (1/k)\log(T/A), \qquad \Omega(T) = A\exp(kT) \tag{11}$$

and

$$\theta(T) = \alpha T, \tag{12}$$

where $\alpha = \exp(k\lambda)$.

It is well known, and easily shown, that the relative motions of the observers in the case $n = 1$ are all uniform. The relative motions in other cases, $n \neq 1$, can also be readily obtained. Regraduating from T to τ by equation (5), and adopting the same invariant value c for the signal velocity both before and after regraduation, we deduce the relation *

$$\Omega^{-1}(T + R/c) - \Omega^{-1}(T - R/c) = 2\lambda, \tag{13}$$

for the distance R of one observer from the other at time T. Applying (8) we find that

$$\frac{1}{(T-R/c)^{n-1}} - \frac{1}{(T+R/c)^{n-1}} = 2k(n-1)\lambda. \tag{14}$$

If V denotes the relative velocity we immediately obtain by differentiation the relation

$$\frac{1+V/c}{1-V/c} = \left(\frac{T+R/c}{T-R/c}\right)^n. \tag{15}$$

Differentiating logarithmically with respect to T, we find that

$$\frac{dV/dT}{1-V^2/c^2} = n\frac{(VT-R)}{(T^2-R^2/c^2)}. \tag{16}$$

Although we have assumed that $n \neq 1$ in deriving this formula, equation (15) and hence equation (16) are also valid when $n = 1$. Moreover, it is of interest to note that when $n = 1$ equation (16) is formally identical with the one-dimensional form of the equation of motion of a *free particle* (not necessarily a fundamental particle) in Milne's substratum (**6**). We shall return to this result when considering the motion of free particles.

An alternative formula to (16) can be derived from (14) in the following way. Writing

$$T_1 = T - R/c, \qquad T_2 = T + R/c,$$

* This relation is derived from the identity $(\tau+\lambda)-(\tau-\lambda) = 2\lambda$, where $c\lambda$ is the (constant) τ-scale distance apart at epoch τ.

we deduce, by differentiating (15) with respect to T, that

$$\frac{dV/dT}{1-V^2/c^2} = \frac{nc}{2}\left(\frac{1+V/c}{T_2} - \frac{1-V/c}{T_1}\right),$$

whence using (15) again we obtain

$$\begin{aligned}\frac{dV/dT}{(1-V^2/c^2)^{3/2}} &= \frac{nc}{2}\left\{\frac{(T_2/T_1)^{n/2}}{T_2} - \frac{(T_1/T_2)^{n/2}}{T_1}\right\}\\ &= \frac{nc}{2}(T_1T_2)^{n/2-1}\left\{\frac{1}{T_1^{n-1}} - \frac{1}{T_2^{n-1}}\right\}\\ &= nc(T_1T_2)^{n/2-1}\{k(n-1)\lambda\}\end{aligned}$$

by equation (14). Consequently,

$$\frac{dV/dT}{(1-V^2/c^2)^{3/2}} = n(n-1)(T^2-R^2/c^2)^{n/2-1}kc\lambda. \tag{17}$$

The expression on the left-hand side, which is the same as

$$\frac{d}{dT}\left\{\frac{V}{(1-V^2/c^2)^{1/2}}\right\},$$

we call the "relativistic acceleration" of the observer, or particle, concerned. We note that for $n=2$ this is a non-zero constant, in agreement with previous work on uniformly accelerated observers (7). When $R \ll cT$, $V \ll c$, equation (17) gives

$$dV/dT \sim n(n-1)kc\lambda T^{n-2},$$

so that the acceleration under these conditions is approximately proportional to the $(n-2)$th power of the epoch.

The critical case dividing those in which the relative motion is one of acceleration from those in which it is one of deceleration arises when $n=1$. In this case the acceleration is zero and the motion is uniform. If $n>1$, the particles coincide at $T=0$, as in the case when $n=1$. This can be seen most directly from formula (10). We have

$$\theta(T) = \frac{T}{\{1-k(n-1)\lambda T^{n-1}\}^{1/(n-1)}}. \tag{18}$$

Consequently $\theta(T)=T$ when $T=0$, i.e. the particles meet at $T=0$ (8). Since in the subsequent motion $R<cT$, it follows that because $n>1$ the acceleration given by (17) will be positive, tending to zero as $V\to c$. From (15) we see that $V\to c$ only as $R\to ct$, i.e. after infinite time. Initially when $T=0$ the relative velocity is zero, for from (14) and (15) we deduce that

$$\frac{V}{c} = \frac{1-\{1-2k(n-1)\lambda T_1^{n-1}\}^{n/(n-1)}}{1+\{1-2k(n-1)\lambda T_1^{n-1}\}^{n/(n-1)}}, \tag{19}$$

where T_1 denotes $(T-R/c)$, and since $n>1$ this gives zero V when $T_1=0$, i.e. at $T=0$.

On the other hand, if $0<n<1$, equation (19) is best written in the form

$$\frac{V}{c} = \frac{\{T_1^{1-n}+2k(1-n)\lambda\}^{n/(1-n)} - T_1^n}{\{T_1^{1-n}+2k(1-n)\lambda\}^{n/(1-n)} + T_1^n}, \tag{20}$$

whence we immediately see that $V\to c$ as $T_1\to 0$, when $R=cT$. From (10) it is clear that $\theta(T_1)\neq T_1$ when $T_1=0$, i.e. the particles are not coincident when $V=c$. For $T_1>0$, i.e. $R<cT$, we see from (17) that the velocity continually decreases,

tending to zero as T_1 tends to infinity. If there is a finite maximum value of λ, then the system could be said to expand from an initial radius

$$\frac{c}{2}\{2k(1-n)\lambda_{\max}\}^{1/(1-n)}.$$

The one value of n in the range $0<n<1$ which we will examine in detail is $n=\frac{2}{3}$. Since

$$T_1^{1-n}=T_2^{1-n}-2k(1-n)\lambda,$$

by repeated application of (10), it follows that

$$\frac{V}{c}=\frac{T_2^n-\{T_2^{1-n}-2k(1-n)\lambda\}^{n/(1-n)}}{T_2^n+\{T_2^{1-n}-2k(1-n)\lambda\}^{n/(1-n)}}. \tag{21}$$

If the life-history of each constituent fundamental "observer" begins at local time zero, the first signal from an observer B related to A by signal function (10) will be received by A at local time

$$T_2=\{k(1-n)\lambda\}^{1/(1-n)}.$$

Then, if $n=\frac{2}{3}$, the velocity which A will regard B as having when this signal left B will be zero. A signal leaving A at time $T_1=0$ will reach B at local time $\{k(1-n)\lambda\}^{1/(1-n)}$ when the velocity of B relative to A will be instantaneously c. After that B will gradually decelerate relative to A.*

Thus, if $n>1$, the model expands from a singular point configuration at epoch zero, the initial velocities all being zero and the subsequent relative motions all being accelerated, the relative velocities tending to c after infinite time. If $n=1$ the model expands uniformly from a point configuration at epoch zero. If $n\geqslant 1$, each observer will assign to the model a radius $R=cT$ at time T. If $n=\frac{2}{3}$, the initial configuration at epoch zero is spatially extended and the initial velocities are all zero. Each relative motion attains the velocity c after a finite time and then gradually decelerates to zero after infinite time.†

3. *Hubble's parameter.*—If we assume that for each model the time T is the uniform time of atomic vibrations in that particular model, then the spectra of sources of radiant energy associated with the fundamental particles of the model will all appear to be shifted to the red according to observers attached to these particles. We consider the case of "near" particles for which $R\ll cT$, where R denotes distance apart at time T. From equation (15) it follows that

$$\frac{1+V/c}{1-V/c}\sim\frac{1+nR/cT}{1-nR/cT}, \tag{22}$$

whence

$$V/c\sim nR/cT,$$

where V denotes relative velocity at time T. If we compare this result with Hubble's empirical law for the nearer nebulae,

$$\delta\lambda/\lambda\sim R/ct, \tag{23}$$

---

* In general, if $0<n<1$, equation (21) will be valid only when $T_2>\{2k(1-n)\}^{1/(1-n)}$, corresponding to $T_1=0$.

† Of course, when $n=\frac{1}{2}$ all relative motions will be decelerations throughout, provided that we follow Milne in restricting all measurements of distance to those which could in principle be derived from signals that are first emitted by the observer (subsequent to his initial epoch) and are then reflected back to him by the observed particle instantaneously on arrival.

where $\delta\lambda/\lambda$ denotes the red-shift, R the distance deduced from apparent magnitude * and $t\sim 2\times 10^9$ years, it follows that if $\delta\lambda/\lambda$ is identified with V/c, then

$$T\sim nt. \tag{24}$$

Consequently, Hubble's parameter for these models varies with epoch, being a rough measure of the epoch if $n=1$ and equal to the epoch multiplied by $(1/n)$ in other cases. Thus, if $n=1$, the "present age" of the model, i.e. its age corresponding to the present empirical value t_0 of t, is approximately 2×10^9 years; if $n<1$ its age is less than this figure, e.g. if $n=\frac{2}{3}$ it is about 1·2 (or 1·3) $\times 10^9$ years; whereas if $n>1$ its age is greater, e.g. if $n=2$ its age is nearly 4×10^9 years.

We have assumed that (22) can be replaced by (23), but as in general the relative motion is not uniform it is desirable to justify this step by an independent argument. From (7) we see that the signal function correlating neighbouring observers (fundamental particles) is

$$\theta(T)=T+\epsilon T^n,$$

whence, approximately, $\theta^{-1}(T)=T-\epsilon T^n$, since ϵ is small.† It follows (9) that the distance apart at time T is given by

$$R\sim\epsilon cT^n.$$

If λ, ν are the wave-length and frequency of radiation emitted by A according to his measurements and λ', ν' are the wave-length and frequency as measured by B of the radiation arriving at B, then

$$\lambda'/\lambda=\nu/\nu'=dT'/dT,$$

where $\nu dT=\nu' dT'$ is the number of photons emitted by A in an interval of time dT in the direction of B, and dT' is the interval according to B during which he received them. Thus, since

$$T'=T+\epsilon T^n,$$

we find that

$$\lambda'/\lambda=1+n\epsilon T^{n-1},$$

i.e. if we write $\lambda'=\lambda+\delta\lambda$, then

$$\delta\lambda/\lambda\sim nR/cT,$$

which is the same as equation (23) under the condition (24).

4. *The characteristic dimensionless invariant.*—We have already remarked that when $n=1$ the second-order differential equation (16) gives the equation of motion

* A referee has kindly pointed out that as the R in Hubble's empirical law (23) is calculated from the apparent magnitude of a nebula it must be independent of the graduation of the observer's clock, whereas the R in our equation (22) is defined in terms of clock readings and so would appear to depend on the observer's clock graduation. No inconsistency, however, need arise. To show this we first note that our argument here refers only to "near" nebulae, so that the empirical R refers to the distance of a near nebula as determined by light received at a particular epoch, namely, the present. We can thus meet the criticism by stipulating that the particular value of our theoretical R for a *near* nebula, as determined by light received at *one particular epoch* (which can be identified with the present), is *independent of time-scale*. To show that this is possible, we first impose the condition that all time-scales under discussion read the same time and run at the same rate at this particular epoch which we call τ_0. As these time-scales are given by equation (9), there are two arbitrary constants k and A at our disposal, and we can choose these so that the above conditions are satisfied. Hence, writing $R=\frac{1}{2}c(\tau_2-\tau_1)$, where $\tau_2=\tau_0$, $\tau_1=\tau_0-\Delta$ and Δ is small, it follows that to the first order of small quantities R is the same as $\frac{1}{2}c\{\Omega(\tau_2)-\Omega(\tau_1)\}$. For $R=\frac{1}{2}c\Delta$, and

$$\tfrac{1}{2}c\{\Omega(\tau_2)-\Omega(\tau_1)\}\sim\tfrac{1}{2}c\Delta\,\Omega'(\tau_0)=\tfrac{1}{2}c\Delta,$$

since $\Omega'(\tau_0)=1$ because we have stipulated that at time τ_0 all clocks run at the same rate.

† The factor k has here been absorbed into ϵ.

of a free particle in Milne's substratum, according to a fundamental observer in line with the particle. Milne himself recognized that his derivation of the latter (**10**) was somewhat involved, as it depended both on arguments relating to more complex types of model and on comparison with formulae associated with Newton's law of gravitation. In considering the present class of models, corresponding to a general value of n, equation (16) emerges automatically as the characteristic second-order differential equation of each model. (Equation (17), on the other hand, involves a parameter λ which does not characterize a particular model but only a pair of constituent fundamental particles.) Equation (15) is a first-order equation, every solution of which corresponds to the relative motion of two fundamental particles. Every solution of (15) is, of course, a solution of (16), but there will be solutions of (16) which are not solutions of (15). Nevertheless, even when R, T and V are not related by (15), equation (16) has the important property of being covariant for all fundamental observers in line with the initial observer. This is readily verified by transforming mathematically to τ-measure according to equation (5). The general first integral of (16) is immediately seen to be given by

$$\frac{1+V/c}{1-V/c} = K\left(\frac{T+R/c}{T-R/c}\right)^n,$$

where K is an arbitrary constant. Writing T_1, T_2 for $T-R/c$ and $T+R/c$, respectively, we have

$$\frac{dT_2}{T_2^n} = K\frac{dT_1}{T_1^n}.$$

Hence, if σ denotes the τ-scale "distance" of a particle at the conventional epoch τ, it follows, by applying equations (5), (6) and (7) and writing τ_1, τ_2 for the values of τ corresponding to T_1 and T_2, that

$$d\tau_2 = Kd\tau_1,$$

i.e. that

$$d(\tau+\sigma/c) = Kd(\tau-\sigma/c),$$

whence

$$d\sigma/d\tau = \text{constant}.$$

Since $d\sigma$ and τ are invariants for all fundamental observers, this equation is clearly covariant for all fundamental observers in line and thus equation (16) must be so too. This equation, therefore, corresponds to the simplest covariant second-order equation of motion of the fundamental particles according to the τ-scale.

Since equation (16) can be regarded as the characteristic equation of the *second order* associated with the model as a whole, it is to be expected that all solutions of this equation can be correlated with some significant property of the model. We shall now *define* the class of motions corresponding to all these solutions as the class of free motions associated with the model, and a particle moving relative to some fundamental observer with any such motion will be called a free particle. This definition of free motions is, in effect, part of our definition of the model.

This procedure may be justified as follows. First, from a mathematical point of view: the class of free particles, or rather of free motions, thus defined is a natural first extension of the class of fundamental particles, or rather of fundamental motions, of which the model is constructed. Second, from a physical point of view: associated with these free motions will be found properties which can be compared with Newtonian gravitation.

Methodologically, the procedure does not appear to be inferior to that of Milne. He introduced the idea of a free particle in the substratum in these words: "Consider a particle (not a fundamental particle) passing through the position $\boldsymbol{P}$ at epoch t with velocity $\boldsymbol{V}$, all as measured by the observer O at the origin. Then this free test-particle will have some particular acceleration $d\boldsymbol{V}/dt$. This acceleration may depend on $\boldsymbol{P}, t, \boldsymbol{V}$ as arguments; and it may also involve the conventional constant c and the coefficient B defining the density in the substratum. The variables $\boldsymbol{P}$, t, $\boldsymbol{V}$ must be considered as capable of independent variation; for through any arbitrary position $\boldsymbol{P}$, at an arbitrary epoch t, we can suppose a free particle, passing, or projected, with an arbitrary velocity $\boldsymbol{V}$. The complete trajectory of the free test-particle should be obtainable by integration of the differential equation obtained by determining $d\boldsymbol{V}/dt$, the instantaneous acceleration as measured by O, as a function of $\boldsymbol{P}$, t and $\boldsymbol{V}$.... The question whether this problem is soluble without physical appeal is a question in pure logic" (11). Thus, Milne implicitly defined the motion of a free particle as a solution of a second-order differential equation associated with his model. Moreover, it is clear that he regarded all solutions of this equation as possible free motions. His equation was derived by postulating that it must be a vector equation which preserves its form under transformation from the coordinates of one fundamental observer to another. We have found that our equation (16), which refers to measures by an observer in line with the moving particle, possesses a similar covariant property. Unfortunately, Milne's method does not yield a unique equation, since an arbitrary function of two variables arises. The reduction of this function to a unique constant value is not possible without appeal to considerations outside "pure logic". The present procedure is more concise and direct and, like Milne's, is ultimately "justified," by its consequences.

We will now consider a particle moving in line with the observer O (associated with some fundamental particle) according to the characteristic equation of motion (16). In particular, we will suppose that at some instant such a particle is instantaneously at rest with respect to O at a point close to O, i. e. at a point distant R from O where $R \ll cT$. The acceleration according to O of this particle at this instant will then be

$$d^2R/dT^2 = -nR/T^2, \tag{25}$$

where we have replaced V by dR/dT. The form of this equation suggests comparison with Newtonian gravitation. If $R=0$, i. e. if the free particle is instantaneously at rest at O, then its acceleration will be zero, in agreement with the condition that the fundamental particle at O is itself a free particle and also with the condition that O is at a centre of spherical symmetry. If R is not zero, then the free particle at the instant considered will, in general, not be at a centre of symmetry, except when $n=0$, i. e. when the model is static. Thus, when $n>0$, the free particle will be accelerated towards O, the magnitude of the acceleration being proportional to R. If, in accordance with Newtonian ideas, we can regard this acceleration as due to the gravitational pull by that part of the model which lies within a sphere of centre O and (small) radius R, equation (25) should be equivalent to the equation

$$d^2R/dT^2 = -(\tfrac{4}{3}\pi G\rho)R, \tag{26}$$

where G is the constant of gravitation and ρ the (local) material density of the model at O. Since, in general, ρ will depend on the epoch, we will refer to ρ and

G in (26) as the values of the local density and constant of gravitation at the particular epoch T considered.

Comparison of (25) and (26) immediately yields the relation

$$G\rho T^2 = (4\pi/3n)^{-1}. \tag{27}$$

Hence, subject to the comparison here made with Newtonian gravitational theory, each model of the class considered can be correlated with a particular dimensionless invariant $G\rho T^2$, the value of which is uniquely determined by the index n. If we assume that the relative motions of neighbouring fundamental particles of the model give rise to Doppler shifts in accordance with equation (23), so that we can introduce a "Hubble parameter" t which is related to the epoch T by equation (24), viz. $T = nt$, it follows that the system is characterized by an invariant relation

$$G\rho t^2 = \left(\frac{4\pi}{3}\right)^{-1} \frac{1}{n}. \tag{28}$$

In principle, each term on the left can be "identified" with an observable parameter, G, ρ and t being the current values (at any epoch) of the constant of gravitation, the local smoothed-out density, and the ratio of distance to Doppler shift, multiplied by c, of nearby nebulae, respectively.

5. *Comparison of ages and densities of different models.*—In classical physics and in Einstein's general theory of relativity, it is postulated that the constant of gravitation G is an absolute constant of nature. If we impose this particular postulate on the system of world-models here considered, we see from (27) that we isolate the particular model for which $\rho T^2 = \text{constant}$. From equation (7) we have

$$\theta(T) = T + \epsilon k T^n$$

as the signal function correlating two equivalent observers attached to neighbouring fundamental particles of the model. It follows immediately that the distance apart of each particle at time T is proportional to T^n. Consequently, the local density of the model ρ varies as T^{-3n}. Hence, if $\rho T^2 = \text{constant}$, it follows that we must have $n = \frac{2}{3}$. We have already remarked that the relative motion for this value of n is, in general, one of deceleration. From (24) we see that the present age of such a system corresponding to the current value of t would be $\frac{2}{3}t \sim 1{\cdot}2 \times 10^9$ years. From (28) we deduce that the corresponding current value of the local mean density ρ should be approximately $1{\cdot}5 \times 10^{-27}$ g./cm.^{3}. This density is higher by a factor $\frac{3}{2}$ than that associated with Milne's model, which corresponds to $n = 1$. The age is comparable with the estimate previously derived by Hubble (**12**) from a theoretical analysis of the empirical data. It is well known that this estimate is difficult to reconcile with other empirical and theoretical requirements and that Hubble's analysis has been subjected to serious criticism (**13**). We conclude, however, that if we adopt:

(i) the classical postulate of the invariance of G with lapse of time;

(ii) the postulate that all spectral shifts of nebulae are due to the Doppler effect;

(iii) a homogeneous world-model of the type considered above for which the product $G\rho t^2$ is an invariant;

then we obtain a value for the local density which is on the high side and a value for the age of the system which appears to be far too low.

In order to obtain more acceptable estimates for the age and density, we must

consider values of n greater than $\frac{2}{3}$. For such values G will no longer be independent of epoch. Indeed, from the invariance of $G\rho T^2$ and the fact that ρ varies as T^{-3n}, we deduce in general that

$$G \propto T^{3n-2}. \tag{29}$$

In particular, if $n = 1$ we see that $G \propto T$, a result originally due to Milne. The total gravitational force exerted by one nebula on a neighbouring one, neglecting any possible variations in their masses, will be proportional to G divided by the square of their distance apart, i. e. to T^{n-2}.

We can now construct the following table for the properties of models corresponding to different values of n. We observe that by appropriate choice of n

TABLE I

n	Type of model	$G\rho t^2$	" Present age "	" Present density "	Variation of G	Gravitational force
$\frac{2}{3}$	Deceleration	$9/8\pi$	$1{\cdot}2 \times 10^9$	$1{\cdot}5 \times 10^{-27}$	Constant	$T^{-4/3}$
1	Uniform expansion	$3/4\pi$	$1{\cdot}9 \times 10^9$	10^{-27}	T	T^{-1}
2	Uniform acceleration	$3/8\pi$	$3{\cdot}7 \times 10^9$	5×10^{-28}	T^4	Constant
3	Increasing acceleration	$1/4\pi$	$5{\cdot}6 \times 10^9$	$3{\cdot}3 \times 10^{-28}$	T^7	T
10	Extremely rapid increasing acceleration	$3/40\pi$	$1{\cdot}9 \times 10^{10}$	10^{-28}	T^{28}	T^8

The " present age " is in years, the " present density " in g./cm.^{3}. The final column gives the variation in time of the total gravitational force exerted by one nebula on a neighbouring one.

we can increase the age of the system so as to equal or exceed any of the current estimates based on considerations of stellar evolution, etc. For example, if $n = 2$, the age corresponding to the present determination of Hubble's parameter is nearly 4×10^9 years. On the other hand, the estimate of the current local density of the model is relatively insensitive to increase of n. For n lying in the range $1 \leqslant n \leqslant 10$, the density lies in the range $10^{-27} \geqslant \rho \geqslant 10^{-28}$, where ρ is calculated in c.g.s. units. Since many independent lines of evidence suggest that the age of the universe is unlikely to be less than about 2×10^9 years, it follows that the probable range in which n lies is bracketed below by the value $n = 1$. A *possible* criterion for determining an upper bracket is suggested by inspection of the last column in Table I. As the nebulae travel farther and farther apart it would seem to be physically implausible for the total gravitational pull of one on the other actually to increase with time. This suggests that $n = 2$ is a reasonable upper bracket *, so that we have

$$1 \leqslant n \leqslant 2,$$

and consequently, if the universe can be compared with models of this class, its age and mean local density at present must have values T years and ρ g./cm.^{3}, respectively, where

$$\sim 2 \times 10^9 \leqslant T \leqslant \sim 4 \times 10^9$$

and

$$10^{-27} \geqslant \rho \geqslant 5 \times 10^{-28}.$$

Incidentally, we note that the gravitational tide-raising force of one nebula on another (neighbouring) nebula will diminish whichever model of the class we choose. For this force will depend on G/r^3, where r is the distance apart of the nebulae concerned, and this expression varies as T^{-2} for all values of n.

* This case gives a model in which all fundamental particles are in uniform relativistic acceleration. The properties of frames of reference in uniform relativistic acceleration have been discussed by many authors (14).

6. *Conclusions.*—The fact that Gpt^2 is a dimensionless quantity with an empirically determined value which is not widely different from unity suggests that it may be an invariant characterizing some property of the physical universe. The invariance of a product of this type is found to be a property of a class of homogeneous world-models of which Milne's uniformly expanding substratum is a particular case.

Milne's uniformly expanding substratum, regarded as a world-model, is based on the following postulates:—

(i) the nebular red-shifts are Doppler shifts due to mutual recession of the nebulae;

(ii) the red-shift of any one nebula as it would appear to a hypothetical observer on any other nebula does not change with lapse of time;

(iii) the system as a whole can be regarded, to a first approximation, as homogeneous and isotropic about each nebula, and its contents are conserved.

Together, the first two postulates yield uniformity of relative motion, and the resulting model has a present age of about 2×10^9 years and a mean local density of about 10^{-27} g./cm.^{3}.

The technique of kinematic relativity can be adapted, however, to models for which postulate (ii) is relaxed, in particular to the class of models here considered for which Gpt^2 is invariant. Lower densities than 10^{-27} g./cm.^{3} and greater ages than 2×10^9 years can be obtained, although it is unlikely that the former can be reduced by more than one-half, and the latter more than doubled. These results are independent of curvature, and so are the same whether the models are open or closed. In so far as discussion of this class of model is relevant to our study of the extragalactic universe, the strongest result which emerges from the present investigation is that the mean density almost certainly exceeds 10^{-28} g./cm.^{3}. This result suggests that current empirically based estimates may be underestimates, due presumably to the neglect of much non-luminous and (possibly) diffuse matter.

Department of Mathematics,
Imperial College of Science and Technology,
London, S.W.7 :
1951 *May* 18.

References

(1) G. J. Whitrow, *Nature*, **158,** 165, 1946.
(2) P. Jordan, *Nature*, **164,** 638, 1949.
(3) E. A. Milne, *Kinematic Relativity*, Oxford, p. 77, 1948.
(4) *Transactions of the International Astronomical Union*, **7,** 283, 1950.
(5) G. J. Whitrow, *Q. J. Math. (Oxford)*, **6,** 253, 1935.
(6) E. A. Milne, *op. cit.*, p. 26.
(7) E. A. Milne and G. J. Whitrow, *Zeits. f. Astrophys.*, **15,** 342, 1938.
(8) E. A. Milne and G. J. Whitrow, *Phil. Mag.*, **40,** 1247, 1949.
(9) G. J. Whitrow, *Q. J. Math. (Oxford)*, **6,** 251, 1935.
(10) E. A. Milne, *Kinematic Relativity*, Oxford, p. 136, 1948.
(11) E. A. Milne, *op. cit.*, pp. 64–5.
(12) E. P. Hubble, *The Observational Approach to Cosmology*, Oxford, p. 62, 1937.
(13) G. C. McVittie, *Proc. Phys. Soc.*, **51,** 529, 1939. O. Heckmann, *Theorien der Kosmologie*, Berlin, p. 65 *et seq.*, 1942.
(14) H. Bateman, *Proc. Lond. Math. Soc.* (2), **7,** 70, 1909. E. Cunningham, *Proc. Lond. Math. Soc.* (2), **8,** 77, 1910. L. Page, *Phys. Rev.*, **49,** 254, 1936. H. P. Robertson, *Phys. Rev.*, **49,** 755, 1936. E. A. Milne and G. J. Whitrow, *Zeits. f. Astrophys.*, **15,** 342, 1938. E. L. Hill, *Phys. Rev.*, **72,** 143, 1947.

ACCRETION AND THE ORIGIN OF COMETS

P. J. D. Gething

(Communicated by G. J. Whitrow)

(Received 1951 May 3)

Summary

The effect of small density asymmetries in an interstellar dust cloud on the process of accretion by the Sun in its passage through the cloud is considered, with particular reference to Lyttleton's theory of the origin of comets. Such variations in density can give the accreted material a velocity component perpendicular to the axis of accretion sufficient to prevent the material falling into the Sun. Nuclei which might form as a result of the disruption of the accreted material into separate segments can be deflected into periodic orbits, thus apparently removing one of the difficulties of Lyttleton's theory. It appears to be unlikely, however, that the steady state considered by Bondi and Hoyle in earlier papers concerning accretion problems would be reached in the presence of asymmetries. Thus their estimates of the rate at which the Sun can gain in mass might need to be considerably reduced. It is also shown that the influence of random magnetic fields on charged particles will tend to reduce the rate of accretion.

---

1. *Introduction.*—Lyttleton (1) has suggested that comets are formed during the passage of the Sun through an interstellar dust cloud. Particles stream towards the axis of the Sun's motion under the action of the solar gravitational field and, provided the probability of collisions between the particles is sufficiently high, an accretion will form on the axis. This will disrupt into comparatively short line elements, each element representing a potential cometary nucleus in Lyttleton's theory. However, such material should fall directly into the Sun and, if the origin of comets is to be explained on this basis, it is necessary to find a mechanism capable of deflecting some of the nuclei (perhaps only a very small fraction of the total number formed) into periodic orbits round the Sun.

Lyttleton suggested two ways in which nuclei might acquire angular momenta of the required order of magnitude:

(i) the deflecting action of the combined gravitational fields of the planets, and

(ii) the deflecting action of the gravitational fields of other stars.

Another possibility which will be examined in this paper is that the angular momentum arises from a lack of symmetry in the cloud. Within the average cloud we should expect the density to decrease gradually from the central region towards the boundary, apart from local fluctuations of a random character. It is most unlikely that the line of the Sun's motion relative to the cloud would be an axis of density symmetry. Hence the number of particles reaching the axis from different directions will be unequal and any accretion which begins to form will acquire a component of momentum perpendicular to the axis. It will be shown, first of all, that this effect is likely to be sufficient to account for the angular momenta of the comets.

The material which can be captured by the accretion is initially contained within a cylindrical region of the cloud with the line of the Sun's motion as axis. According to Lyttleton (1), a particle, whose undisturbed position was at a perpendicular distance p from the axis, crosses the axis at a distance x from the Sun given by

$$x = \frac{p^2V^2}{2GM},$$

where V is the velocity of the Sun through the cloud, G is the gravitational constant and M is the mass of the Sun. Hence a length dx on the axis corresponds to an *initial* difference in distance from the axis of amount

$$dp = \frac{dx}{V}\sqrt{\left(\frac{GM}{2x}\right)},$$

and the material reaching the element dx throughout the Sun's passage through the cloud will have been drawn from the volume between two circular cylinders with the line of the Sun's motion as common axis and radii p and $p+dp$, respectively. The material arriving from different directions in a given time interval will, therefore, have come from different elements of the cloud of maximum separation $2p$. Between such points, e.g. A and B in Fig. 1, the cloud may exhibit a gradual change in density, such that the density near A is

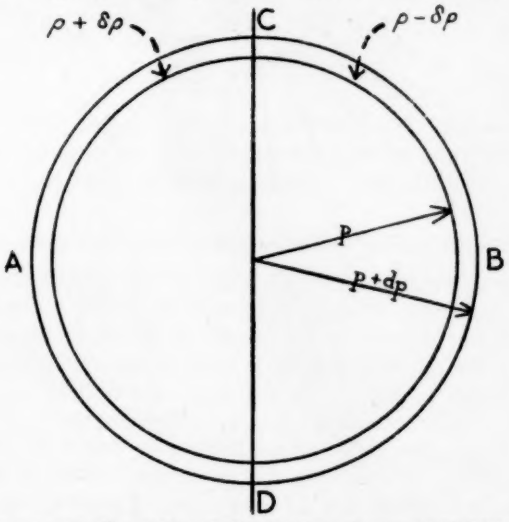


FIG. 1.—*Section of the cylindrical volume element, divided into two regions of uniform but unequal density.*

$\rho+\delta\rho$ and near B it is $\rho-\delta\rho$. A highly idealized model of the cloud will be considered, in which a plane CD through the axis is supposed to divide the cylindrical region into two equal volumes of uniform but unequal density. On one side of the plane the density is everywhere $\rho+\delta\rho$ and on the other it is $\rho-\delta\rho$; it is to be expected that the degree of asymmetry, $\delta\rho/\rho$, will increase with increasing radius p of the cylinder.

In general, of course, the density variations within the cloud would probably be gradual, and such a discontinuity would not exist. However, in the absence of precise information about the density variations within the cloud under

consideration, there seems to be little point in assuming a particular angular distribution of the excess particles. The uniform distribution shown in Fig. 2, and the implied density model, as shown in Fig. 1, will therefore be adopted for mathematical simplicity; the choice of angular distribution has very little effect on the magnitude of the angular momentum acquired by accreted material in the manner suggested above. The values for the densities in the two halves of the cylindrical region may be assumed to represent average densities in the two volumes and, provided $\delta\rho/\rho$ is small, the results obtained from a consideration of the discontinuous model should give a good approximation to the results to be expected in practice.

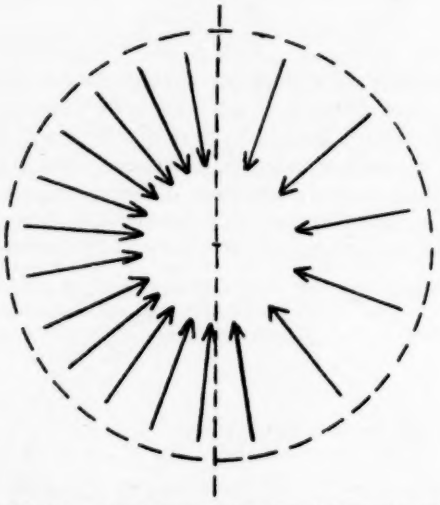


FIG. 2.—*The distribution of arriving particles adopted for mathematical simplicity.*

It will be further assumed that the value of $\delta\rho/\rho$ appropriate to a particular value of the axial distance x remains constant with time, i.e. that the degree of asymmetry does not depend on the depth of the Sun's penetration into the cloud. Whereas the previous restriction on $\delta\rho/\rho$, apparently implying a density discontinuity, could be easily removed by a re-definition of the "degree of asymmetry", this second restriction is a physical assumption as to the distribution of the mass within the cloud which is not likely to be strictly satisfied in a practical case. The possibility of removing the restriction is considered at the end of Section 5, where it is shown that the model here adopted introduces the feature of asymmetry in the only form in which it can be simply treated. Other models exhibiting a continuous variation in density would lead to a gradually changing velocity component due to asymmetry of the line accretion, the magnitude of this acceleration depending on the mass of the accretion. This would make the problem less tractable since, in clouds of low density, the process of accretion is likely to be very slow in starting, and no reliable estimate could be made of the line density of the accretion at a given time after the Sun's entry into the cloud.

2. *Magnitude of the angular momentum arising from lack of symmetry.*—In the symmetrical model originally considered by Hoyle and Lyttleton (**2**), the mass of material reaching the axis per unit length per unit time is

$$m = 2\pi\rho_\infty GM/V,$$

where ρ_∞ is the density of the cloud at very large distances from the Sun. For small asymmetries a similar formula will apply, with mass $\pi(\rho-\delta\rho)GM/V$ arriving from the low density region and $\pi(\rho+\delta\rho)GM/V$ from the high density region. The excess mass arriving from the high density side will then be

$$2\pi GM\delta\rho/V,$$

representing a fraction $\delta\rho/\rho$ of the total mass arriving in the same time interval.

A particle which crosses the axis at a distance x from the Sun has a velocity component perpendicular to the axis of $\sqrt{(2GM/x)}$, so that the average velocity component perpendicular to the dividing plane for all the particles arriving from one side of the plane at this distance is $\frac{2}{\pi}\sqrt{\left(\frac{2GM}{x}\right)}$. The momentum perpendicular to the plane contributed by the particles from the high density side will exceed that of the particles from the low density side by the amount

$$\frac{4GM\delta\rho}{V}\sqrt{\left(\frac{2GM}{x}\right)}.$$

An accretion which starts to form on the axis will be pushed towards the low density side by the excess momentum, moving with a velocity U such that the momenta of arriving particles relative to the moving aggregate will just cancel to zero, i.e. such that

$$\frac{\pi GM}{V}(\rho+\delta\rho)\left\{\frac{2}{\pi}\sqrt{\left(\frac{2GM}{x}\right)}-U\right\}=\frac{\pi GM}{V}(\rho-\delta\rho)\left\{\frac{2}{\pi}\sqrt{\left(\frac{2GM}{x}\right)}+U\right\},$$

giving

$$U=\frac{2}{\pi}\sqrt{\left(\frac{2GM}{x}\right)}\frac{\delta\rho}{\rho}. \qquad (1)$$

It was shown by Lyttleton that x, the distance of a nucleus from the Sun, must lie between about 17 and 1100 astronomical units. Substituting the larger value (i.e. $1{\cdot}64\times10^{16}$ cm.) in order to find a lower limit for the asymmetry velocity, we obtain

$$U=(8{\cdot}1\times10^{4})\delta\rho/\rho \text{ cm. sec.}^{-1}.$$

Thus for an asymmetry as small as one part in a hundred the transverse velocity of the most distant segments of the accretion would be about 10^3 cm. sec.$^{-1}$. The velocities of segments nearer the Sun, with the same value of $\delta\rho/\rho$, could be up to about seven times greater, but it must be remembered that small values of x correspond to small values of p, the initial perpendicular distance from the axis, and the value of $\delta\rho/\rho$ over a limited region of the cloud might be rather less than the value corresponding to a larger value of x.

It will be assumed that at large distances the accreted material resulting from many collisions between arriving particles is moving neither towards nor away from the Sun; this condition, in fact, determines the upper limit for x, giving the so-called "neutral point". Thus for x near $1{\cdot}64\times10^{16}$ cm. we shall assume that the radial velocity of the accreted material is zero, or so small that it may be ignored. Then h, the angular momentum per unit mass of the nucleus with respect to the Sun, will be Ux, i.e.

$$h=\left\{\frac{2}{\pi}\sqrt{(2GMx)}\right\}\frac{\delta\rho}{\rho}. \qquad (2)$$

With the typical values used previously, i.e. $x=1{\cdot}64\times10^{16}$ cm. and $\delta\rho/\rho=0{\cdot}01$, h is found to be $\sim10^{19}$ cm.^{2} sec.$^{-1}$.

3. *Gravitational pull of the cloud.*—In general the cloud will exert a gravitational pull on the nucleus which might tend to reduce the asymmetry

velocity. However, the Sun and the particles streaming towards the axis will also be subject to similar forces, which would be expected to be directed towards the high density region of the cloud. Thus it is the difference between the gravitational attraction per unit mass on the Sun and that on the nucleus which determines the effect on the angular momentum. It will now be shown that this difference is likely to be so small that its influence can be entirely neglected.

Consider the gravitational force per unit mass on a body at a distance r from the centre of a spherically symmetrical cloud, in which the density is a function of r only. The force is $4\pi Gr\rho_r/3$, where ρ_r is the average density of the material contained in a sphere of radius r. For ρ_r we insert the value used for the cloud density in Lyttleton's paper, namely 10^{-24} g. cm.$^{-3}$ and, with $r \sim 0{\cdot}1$ light-years, the force proves to be $\sim 2{\cdot}6 \times 10^{-14}$ dynes/g.

In a cloud of more irregular shape and density distribution somewhat larger forces might arise. On the other hand, we are only interested in the differences in the gravitational field which could arise within a comparatively small region of the cloud, so that 10^{-13} dynes g.$^{-1}$ would appear to be a generous estimate of the differential force per unit mass.

The direction of the force is unimportant, since it is obvious that, even if it were directly opposed to the drift velocity, its effect would be entirely negligible. For example, the asymmetry velocity $\sim 10^3$ cm. sec.$^{-1}$ previously calculated for $\delta\rho/\rho = 0{\cdot}01$ would be reduced by only about 0·01 per cent during the total time of the Sun's passage through the cloud. The effect on the velocities of dust particles streaming towards the accretion axis would be even smaller, since the velocities are about one hundred times larger than that of the aggregate and the period during which the force is effective is much shorter.

4. *Comparison with observation: angular momenta of periodic comets.*—The angular momentum per unit mass, h, can be readily estimated for those comets known to be periodic. In terms of the semi-major axis a and the eccentricity e, the theory of orbits gives

$$h = \sqrt{\{GMa(1-e^2)\}}.$$

Comparison with expression (2) gives

$$\frac{\delta\rho}{\rho} = \frac{\pi}{2}\sqrt{\left\{\frac{a(1-e^2)}{2x}\right\}}. \qquad (3)$$

Thus the values of $\delta\rho/\rho$ required to account for the angular momenta of comets for which a and e are known can be calculated. With $x = 1{\cdot}64 \times 10^{16}$ cm. the values required for the short-period comets are about 0·03. As specific examples, Comet Encke (period 3·297 years) would require $\delta\rho/\rho = 0{\cdot}027$ and Halley's comet (period 76·03 years) would require $\delta\rho/\rho = 0{\cdot}037$. Variations in density of the order of a few parts in a hundred should be common within a cylindrical region of the cloud of radius 100 astronomical units (say) from which the accretion material is drawn. If the cloud has a fairly sharply defined boundary there is no reason why the value of $\delta\rho/\rho$ at the boundary should not approach unity.

For comets with larger orbits, which may exist in vast numbers (see (3)), the values of h would be much greater and, since $\delta\rho/\rho$ must be less than unity, it may prove impossible to account for the momenta of such comets by the theory developed above. This is a difficulty inherent in Lyttleton's original suggestion of origin by accretion; it is difficult to see how, on the basis of his theory, the value of a in expression (3) could become greater than x, the distance of formation. It is interesting to note that if we specify $a \leqslant x$ and e close to unity, the value of

$\delta\rho/\rho$ demanded by (3) is automatically less than one. This is very satisfactory, since it appears that small asymmetries in density are capable of providing nuclei with angular momenta of the right order in just those cases to which Lyttleton's theory is applicable.

5. *Effect of asymmetries on "Steady State" conditions.*—Bondi and Hoyle (**4**) considered the possibility of the axial condensation reaching a steady state, in which the mass of material arriving at a given line element per unit time is just equal to the mass of material streaming from the element along the axis (and eventually falling into the Sun) in the same period. Thus the line density remains constant, and is of the order $2\pi\rho G^2M^2/V^4$.

The minimum time required to reach the steady state under idealized conditions of perfect symmetry may be readily estimated. Material crosses unit length of the accretion axis at the rate of $2\pi\rho GM/V$ g. sec.$^{-1}$ cm.$^{-1}$, so that even if all the material were captured the minimum time required to reach the line density quoted above would be

$$t = GM/V^3 \text{ sec.},$$

ignoring the loss of mass by streaming along the axis during the time of formation. For the solar speed of 1 km. sec.$^{-1}$ used by Lyttleton, t proves to be about 4200 years.

However, if cometary nuclei can acquire sufficient angular momentum to give them periods of only a few years by the asymmetry mechanism considered above, it would appear that the assumption that the condensation remains on the axis throughout the passage of the Sun through the cloud, taking perhaps 10^5 years, cannot be justified in the type of model considered here. The question then arises as to whether the steady state can ever be reached.

There are two distinct factors which might tend to reduce the rate at which the Sun can gain in mass:

(i) the possible deflection of accreted material into periodic orbits by the asymmetry mechanism considered above, thus preventing its capture by the Sun;

(ii) the sweeping of material from the axis, which might reduce the probability of an arriving particle being captured by collision with the axial accretion.

So far as the first factor is concerned, there will clearly be a critical value of the asymmetry $\delta\rho/\rho$ above which the asymmetry velocity is sufficient to prevent material falling into the Sun. This value will now be calculated, on the assumption that the accreted material travels in an elliptic orbit round the Sun, passing through the point of formation at distance x in a direction perpendicular to the axis with angular momentum h given by equation (2). If the material is not to fall into the Sun, then its distance from the centre of the Sun at the point of closest approach must clearly exceed one solar radius, i.e.

$$a(1-e) > r_\odot,$$

where $r_\odot$ is the radius of the Sun. This condition, together with the equation for the distance of formation,

$$x = a(1+e),$$

and a modified form of equation (3),

$$\left(\frac{\delta\rho}{\rho}\right)^2 = \frac{a\pi^2(1-e^2)}{8x},$$

yields the required restriction on the degree of asymmetry,

$$\frac{\delta\rho}{\rho} > \frac{\pi}{2}\sqrt{\left(\frac{r_\odot}{x-r_\odot}\right)}.$$

For example, material captured at the axis at a distance $x = 10^{16}$ cm. will not approach nearer than one solar radius to the Sun's centre provided

$$\frac{\delta\rho}{\rho} > 0{\cdot}004.$$

As pointed out in Section 4, asymmetries much greater than this might be expected in practice; indeed, Hoyle and Lyttleton (**5**) themselves mention the probable existence of variations in density within a cloud by factors of a few powers of ten. Thus we see that the rate at which the Sun can gain in mass by accretion is very sensitive to the existence of minute asymmetries. The total gain in mass of the Sun during its passage through the cloud might, in practice, be negligible compared with the theoretical estimate given by Bondi and Hoyle (**4**), on which many far-reaching theories have been based.

With regard to the second factor, it seems unlikely that the loss of target particles due to their diffusion from the axis would have any appreciable effect on the rate of capture of arriving particles. A particle approaching the axis may suffer collisions which can be divided into two main types, either (*a*) with another incoming particle of approximately the same mass, or (*b*) with a rather more massive particle moving slowly in the swarm of accreted particles. The loss of accreted particles from the axial region by transverse asymmetry streaming may reduce the number of collisions of type (*b*), but it cannot affect the frequency of type (*a*) collisions. The problem considered in this paper, viz. the formation of accretion nuclei possessing velocity components perpendicular to the axis, could only arise provided the incoming particles suffer many mutual collisions near the axis in the initial stages of formation of the accretion, losing nearly all their original momenta and retaining only the common component due to the density asymmetry. Hence we have implicitly assumed that collisions of type (*a*) are so common that any particle reaching the axis is almost certain to be captured, so that the fraction captured of the mass arriving at the axis in a given time interval is always very close to unity whether collisions of type (*b*) are frequent or not. It would therefore appear that, provided the cloud is sufficiently dense for an accretion to begin to form, factor (ii) is not likely to reduce the rate of capture of mass at the axis appreciably.

Mutual collisions are less certain to occur initially in a cloud of density less than about 10^{-24} g. cm.$^{-3}$, and the possibility of the formation of an accretion would then require a more detailed investigation than has yet been attempted. It seems unlikely that the existence of a small density asymmetry would have any significant effect on the rate of accretion in such circumstances; the basic problem would be to find the fraction α of arriving particles captured at the axis as a function of time, $\alpha = \alpha(t)$.

It has been pointed out that the degree of asymmetry, $\delta\rho/\rho$, is also likely to vary with the depth of the Sun's penetration into the cloud, i.e. with time. In conjunction with the possible variation of $\alpha(t)$, this suggests four possible cases for theoretical investigation:

(i) $\delta\rho/\rho$ constant and $\alpha(t)$ always very close to 1.
(ii) $\delta\rho/\rho$ constant and $\alpha(t)$ varying with t,
(iii) $\delta\rho/\rho$ varying, $\alpha(t)$ always very close to 1,
(iv) both $\delta\rho/\rho$ and $\alpha(t)$ varying with t.

The approximate calculations given above, relating to the angular momenta acquired by nuclei and the minimum value of $\delta\rho/\rho$ which prevents material captured at a given distance along the axis from falling into the Sun, apply to cases (i) and (ii). An analysis of case (iii) would probably not present any mathematical difficulty; it has not been attempted here, however, because observational data are at present insufficient to suggest a detailed model of density variations. Similarly, an investigation of case (iv) would also require extensive data relating to density variations within the particular cloud under consideration, although in this case the analysis might be much more complex.

Thus we see that the rate of solar accretion may be very sensitive to the particular path followed by the Sun in its passage through the cloud and to the density distribution within the cloud. Any attempt to give a general estimate of the rate of accretion based on an idealized model of an "average" cloud is likely to be of limited value; the results obtained from a consideration of models (i) and (ii), however, give an indication of the importance of asymmetries in considering both the rate of accretion and the origin of comets.

6. *Reconsideration of Lyttleton's theory.*—There seems to be no reason why Lyttleton's theory of the origin of comets, which has many attractive features, should not be retained in a somewhat modified form. His method of estimating the size and mass of a nucleus formed by the disruption of the line filament on the axis is no longer valid in the presence of asymmetries; it seems likely that these quantities might depend on the different transverse velocities acquired by sections of the filament at different distances x along the axis.

The mass of a nucleus would therefore appear to depend on random conditions within the particular cloud under consideration, rather than on general properties common to all clouds. However, the following order-of-magnitude calculation shows that if only a very small number of the particles crossing the axis form into nuclei moving in periodic orbits, the mass of a nucleus might nevertheless be comparable with that of a comet.

In a cloud of average density 10^{-24} g. cm.$^{-3}$, material is crossing each centimetre of the axis at the rate of $2\pi\rho GM/V = 8 \times 10^{-3}$ g. sec.$^{-1}$, with $V = 10^5$ cm. sec.$^{-1}$ as before, so that the total amount crossing the axis between the limits given by Lyttleton, namely about 17 and 1100 astronomical units from the Sun, is $\sim 10^{14}$ g. sec.$^{-1}$. The time of passage through a cloud 0·1 parsec in depth is 3×10^{12} sec., so that the total material crossing the axis would be $\sim 3 \times 10^{26}$ g.

Estimates of the masses of observed comets, which are subject to large errors, range from 10^{16} to 10^{22} g. The material spread along the axis between the limits quoted above would be expected to disrupt into many nuclei, of which only the most distant could escape capture by the Sun; even so, the deflection into a periodic orbit of only one particle in every million crossing the axis might be sufficient to account for the formation of short-period comets.

7. *Influence of random magnetic fields.*—There is another factor which may modify the orbits of dust particles in the solar gravitational field, namely the influence of possible magnetic fields on charged particles. The importance of electric and magnetic fields in cosmical problems has attracted considerable attention since the original papers on accretion were published; the influence of magnetic fields on the accretion process deserves further investigation. Only a brief order-of-magnitude calculation will be given here.

We must estimate the mass and charge of the average dust particle and the magnitude of the fields likely to be encountered within a cloud.

Lyttleton assumed that the linear dimensions of dust particles were $\sim 10^{-7}$ cm. and the density of the material to be about 3 g. cm.$^{-3}$. Hence the mass of a typical particle would be $\sim 10^{-20}$ g.

The average charge carried by dust particles is believed to be about 100 electrons (see (**6**)).

Provided that a very small magnetic field exists in the neighbourhood of the cloud, the motion of the cloud particles relative to the Sun would give rise to a much stronger magneto-hydrodynamic field relative to the Sun in the manner investigated by Alfvén (**7**). The intensity H of this field is given by

$$H = V(4\pi\rho)^{1/2},$$

where ρ is the density of the cloud and V its velocity relative to the observer. For $V = 10^5$ cm. sec.$^{-1}$ and $\rho = 10^{-24}$ g. cm.$^{-3}$, H is found to be $\sim 3{\cdot}5 \times 10^{-7}$ gauss.

To obtain an indication of the influence of this field on the motion of dust particles, consider the radius of the circular orbit of a particle moving in a uniform field of the above intensity with a velocity of one kilometre per second: it is $\sim 10^9$ cm. Since the linear dimensions of the gravitational orbit of a particle in its journey to the accretion axis are $\sim 10^{16}$ cm., the magnetic field is obviously of sufficient intensity to modify the orbits considerably.

Qualitatively, there seems to be no reason why the particle orbits should still pass through the axis of the Sun's motion, or even close to it. In the presence of a force $\boldsymbol{F}$ and a magnetic field $\boldsymbol{H}$, the drift motion of a charged particle is in the plane through $\boldsymbol{H}$ perpendicular to the plane containing $\boldsymbol{H}$ and $\boldsymbol{F}$. When $\boldsymbol{H}$ and $\boldsymbol{F}$ are functions of position, the determination of the motion is a complicated problem, made more difficult by another factor* which must be taken into account, namely the secondary magnetic fields arising from the motions of the charged particles.

It is true that according to some theories (see (**8**) and (**9**)) magnetic fields "regular" in shape will tend to concentrate charged particles into certain regions of the field, possibly giving rise to accretions; the process is helped by the small spiralling motions superimposed on the drift velocities, which increase the probability of collisions. There is no reason, therefore, for supposing that the existence of magnetic fields will make it quite impossible for accretions to form. However, the formation of an axial accretion in the manner envisaged by Hoyle and his collaborators appears to be improbable in such circumstances.

If the cloud contained uncharged particles, their motions would not, of course, be affected by a magnetic field and their orbits would pass through the axis, but the effective density of the cloud (i. e. the density of uncharged particles) would presumably be much smaller than the total density. Reliable observational data on the charges carried by dust grains is obviously required before the rate of growth of an axial accretion can be accurately estimated.

8. *Conclusions.*—A particularly simple model of density variation within an interstellar dust cloud has been considered in investigating the influence of asymmetries in density on accretion. It is found that differences in density of the order of a few parts in a hundred within the tunnel of accretion would be sufficient to give the more distant nuclei of accreted material angular momenta

*A further factor which might have to be taken into account for particles starting from initial positions less than a few astronomical units from the Sun is the Sun's general magnetic field, but according to the latest observations the magnitude of this field is thought to be negligible.

relative to the Sun comparable with the momenta of the periodic comets. Even smaller asymmetries could prevent accreted material falling into the Sun.

The present paper gives rise to considerable doubts as to whether, in general, the accretion process could reach significant proportions in practice. Three distinct factors have been considered which might drastically reduce the rate of gain of mass by the Sun:

(i) the constant sweeping of material from the axis by the asymmetry mechanism, which might reduce the probability of a particle being captured by collision with the axial condensation,

(ii) the possible deflection of accreted nuclei into periodic orbits, preventing the capture of the material by the Sun,

(iii) random magnetic fields existing within the cloud, which will deflect charged dust grains so that their orbits in the solar gravitational field no longer pass through the line of the Sun's motion.

It is suggested that, provided the cloud is sufficiently dense for collisions between pairs of particles approaching the axis to be frequent, the first factor is likely to be relatively unimportant, but that the second and third factors might reduce drastically the rate of solar accretion. Those theories concerning problems of stellar energy generation, climatic variation, etc., which depend on theoretical estimates of the rate at which the Sun can gain in mass would appear to require re-examination.

It is a pleasure to express my thanks to Dr G. J. Whitrow, who directed my attention to this subject, for many helpful discussions and suggestions, and also to Professor W. H. McCrea, who kindly commented on a first draft of the manuscript. I am also grateful to the Department of Scientific and Industrial Research for providing a maintenance grant.

Department of Mathematics,
Imperial College of Science,
London, S.W.7:
1951 *May* 3.

References

(1) R. A. Lyttleton, *M.N.*, **108**, 465, 1948.
(2) F. Hoyle and R. A. Lyttleton, *Proc. Camb. Phil. Soc.*, **35**, 405, 1939.
(3) J. H. Oort, *B.A.N.*, **11**, 91, 1950.
(4) H. Bondi and F. Hoyle, *M.N.*, **104**, 273, 1944.
(5) F. Hoyle and R. A. Lyttleton, *Proc. Camb. Phil. Soc.*, **35**, 599, 1939.
(6) L. Spitzer, *Phys. Rev.*, **70**, 777, 1946.
(7) H. Alfvén, *Cosmical Electrodynamics*, chap. IV, Oxford University Press, 1950.
(8) H. Alfvén, *Stockholms Observatoriums Annaler*, **14**, No. 2, 1942; No. 5, 1943; No. 9, 1946.
(9) H. W. Babcock, *Nature*, **166**, 249, 1950.

THE BLUE SUN OF 1950 SEPTEMBER

R. Wilson

(Communicated by the Astronomer Royal for Scotland)
(Received 1951 June 13)

Summary

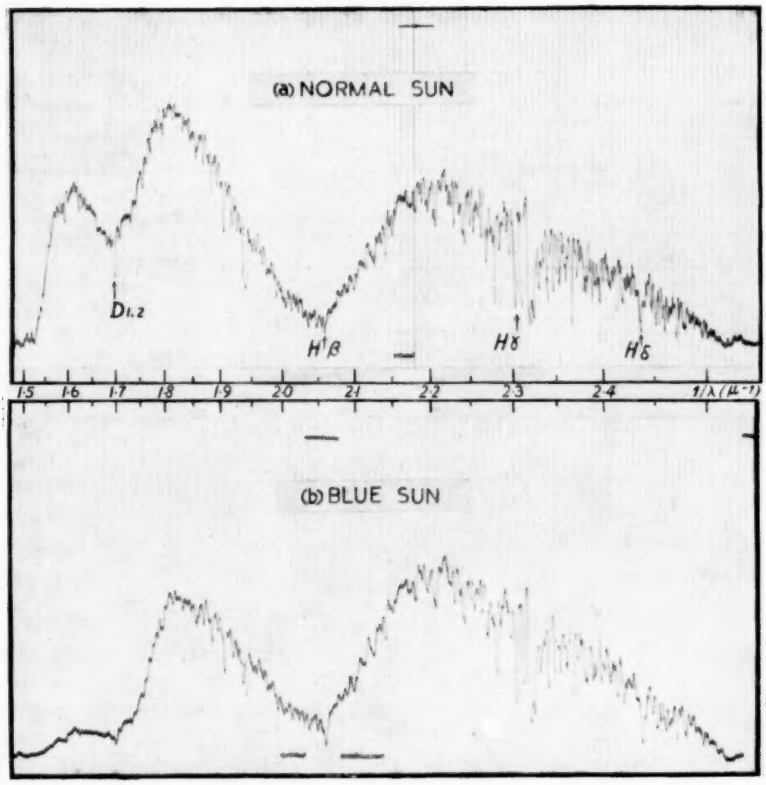
This paper deals with the "blue" Sun observed at Edinburgh on 1950 September 26. Spectrophotometric observations of the "blue" and normal Suns give an extinction curve of the layer causing the phenomenon. This covers the wave-length range 3800–6300 A. and shows blueing on the long wave-length side, and reddening on the short wave-length side, of a minimum situated at λ4350. This affords an experimental verification of an outstanding feature of scattering theory. It is concluded that the scattering particles producing the minimum are dielectric, and uniform in size and constitution. A relation between the size and refractive index of the dielectric particles is deduced. In addition to the selective effect, there is a heavy non-selective extinction which may arise from another system of absorbing particles in the layer. Other workers have established the source of the layer as the extensive forest fires burning in Alberta on 1950 September 23.

---

1. *Observational.*—In Edinburgh on the afternoon of 1950 September 26, the Sun was observed to be a deep indigo blue. This remarkable phenomenon was first perceived at about $15^h 00^m$ U.T., when a thinning out of cloud enabled the solar disk to become visible. From then it could be seen through the cloud for considerable periods up to sunset. The phenomenon continued into the night when a "Blue Moon" was also observed. Next morning, the Sun was normal.

During the blueing period three spectrograms of varying density (exposure times, 10, 30 and 180 seconds) were secured of the Sun's centre on a Kodak P.25 panchromatic plate. The 36-inch Cassegrain reflector of the Royal Observatory was used with the two-glass prism train of the attached Hilger universal spectrograph, giving the respective dispersions 132, 45 and 28 A./mm. at $H\alpha$, $H\beta$ and $H\gamma$. The plate was obtained between $15^h 20^m$ and $15^h 25^m$ U.T., at a solar altitude of 19°. The author was extremely fortunate to obtain this plate. It was $15^h 00^m$ U.T. when the phenomenon was observed, and by $15^h 15^m$ U.T. the instrument was ready for use and trained on the Sun. At $15^h 20^m$ U.T. there was a gap in the cloud, and the above plate was taken as an exposure test. At the end of the 15 minutes required for development and fixing, the cloud had again closed and no more exposures were possible. The spectrophotometric procedure normal to this observatory was not therefore carried out. The emulsion was not calibrated on the same plate as the spectrum to be investigated and the plate was not "brush" developed. Through force of habit, however, the author took the temperature of the developer as well as the duration of development, so that the development of separate calibration plates could be standardized as well as possible.

It was now necessary to obtain an equal altitude comparison with the normal Sun. The difficulty immediately arose as to how to bridge the large intensity



Microphotometer tracings of (a) normal and (b) "blue" solar spectrograms, obtained with 36-inch reflector and attached spectrograph.
(a) Exposure time 30 seconds: aperture $\frac{1}{24}$ square inch.
(b) Exposure time 10 seconds: full aperture.

range of the two sources being compared (a ratio of about 10,000/1). It was decided to stop down the telescope aperture and use a single pencil passing through a mean thickness of the prisms in the spectrograph. This was achieved by means of a cardboard disk into which a hole of the requisite area could be cut. The same three exposure times used for the "blue" Sun spectrograms were adopted, and two sets of normal solar spectrograms of the centre of the disk were obtained, each on a single plate, on the dates 1950 October 26 and 1950 November 15. The respective solar altitudes were 21° and 17° and the respective apertures used were $\frac{1}{24}$ and $\frac{1}{8}$ square inch. These apertures were adopted to give "good" spectrograms for the respective exposure times 30 seconds and 10 seconds. It is interesting to note that the greater difficulties were experienced in obtaining the normal solar spectrograms. The delay was caused by the scarcity of good cloudless skies and by the accuracy required in the cutting of such small apertures to give good densities on high contrast plates. Unlike the case of the "blue" Sun, there were many failures.

Microphotometer tracings of the 30 seconds exposure of the weaker set of normal Sun spectrograms, and the 10 seconds exposure of the "blue" Sun spectrograms, are reproduced in Plate 8. These indicate clearly the extinction of red light in the case of the "blue" Sun.

The emulsion was calibrated on three separate plates, for each of the three exposure times involved, by two exposures in a multiple slit spectrograph. It was considered desirable to measure the denser spectrograms in order to extend the results into the relatively insensitive region of the Hilger spectrograph beyond $1/\lambda = 2{\cdot}47\mu^{-1}$ where the instrumental absorption is heavy (see Plate 8). This necessitated a brighter source in the calibration spectrograph, which is normally used for stellar spectroscopy work involving exposures of about one hour. This was accomplished by placing an extra projector lamp in the calibrator source and reducing the diffusing screen to a filter paper.

Unfortunately, all of the six plates involved in the investigation could not be taken from the same box. Two of these (the calibration plates for the 10 seconds and 180 seconds exposure times) had to be taken from a new box of a different batch. A photometric comparison of the two batches showed reasonable agreement, however, as can be seen from Table I. This gives the values of the contrast factor f, determined from two separate plates taken from the different batches, for various values of the function Δ and the wave-length λ, where

$$f = d\Delta/d(2{\cdot}5 \log_{10} I)$$

$$\text{and} \quad \Delta = \log_{10}\{(1 - T)/T\}.$$

I is the light intensity at λ and T the transparency of the negative, defined as the ratio of the galvanometer deflections observed with the microphotometer for the photographic deposit and the clear plate. The values of f for the two batches are paired together, the first figure referring to the earlier batch.

TABLE I

Values of contrast factor f for different values of Δ *and* λ, *for the two batches of plates used in the present work*

Δ	−1·0	−0·5	0·0	0·5	1·0
λ 6340	1·57 ; ...	1·38 ; 1·35	1·34 ; 1·34	1·57 ; 1·52	2·15 ; 1·86
λ 5250	1·49 ; ...	1·64 ; 1·49	1·71 ; 1·58	1·76 ; 1·61	1·78 ; 1·64
λ 4630	1·54 ; 1·40	1·43 ; 1·36	1·43 ; 1·40	1·51 ; 1·48	1·63 ; 1·60
λ 4210	1·42 ; 1·34	1·34 ; 1·24	1·30 ; 1·20	1·31 ; 1·20	1·37 ; 1·25

It should be noted that the conditions of observation involving the photographing of the spectra of the "blue" and normal Suns on different plates, at different times and with different optical trains, do not favour spectrophotometry of the highest precision. It so happens, however, that the main conclusions of this paper represent very large effects (as will be evident later), and that they cannot be materially affected by the enhanced photometric errors which necessarily arise in an unpredicted "snap" observation, for which no preparation could be made.

The solar and calibration spectrograms were photometered at a number of wave-lengths chosen to avoid deep lines in the solar spectrum. A non-recording microphotometer of Moll type was used with an analysing slit of 25 μ, and each measurement was the mean of ten readings 25 μ apart. This instrument was used in preference to a recording microphotometer, because of its greater flexibility in measuring high densities. Wave-lengths were determined from an identification of solar lines.

The normal and "blue" solar spectrograms with the same exposure times were paired together, and the measurements reduced to the quantity

$$y = 2{\cdot}5 \log_{10}(cI_N/I_B),$$

where I_N and I_B are the respective intensities at wave-length λ of the normal and "blue" Suns. The constant c is the ratio of the apertures used for the "blue" and normal Suns. The factor 2·5 gives a "magnitude scale" which possesses the convenient property that one unit in the second place of decimals is reasonably related to the observational error. The results are given in Table II. Column 1 gives the wave-length in angstroms. Column 2 gives the reciprocal of the wave-length in micron^{-1}. Columns 3, 4 and 5 give the respective values of $2{\cdot}5 \log_{10}(c_1 I_N/I_B)$ for the exposure times 10, 30 and 180 seconds, determined from the combination of the "blue" solar spectra with the denser set of normal solar spectra. Column 6 gives the values of $2{\cdot}5 \log_{10}(c_2 I_N/I_B)$ determined from the combination of the 30 second "blue" solar spectrum with the 30 second exposure of the weaker set of normal solar spectra. These represent all the measurements possible. There are six possible pairs of spectra, but only these four were measurable in each case at corresponding wave-lengths. The constants c_1 and c_2 were determined from the values of the telescope aperture used in each case and given above, and the results combined in the form $2{\cdot}5 \log_{10}(I_N/I_B)$ in column 8.

A comparison of columns 3 and 6 gives an idea of the accuracy of the results. These two sets of values are determined from different "blue" and normal solar spectra and also different calibration plates. Further, one of the calibration plates (used for the values in column 3) was taken from a separate batch, as was mentioned above, and we have a further check on the agreement of the two batches. Since there is a systematic variation in the differences of the values in these two columns, the random error derived from them will have little significance and the variation is expressed in column 7 in terms of α, where α represents the departure of the difference of columns 3–6 from the mean. The agreement shown by the figures is regarded as satisfactory.

The combined results of column 8, representing the extinction of the absorbing layer causing the "blue" Sun, are plotted against $1/\lambda$ (μ^{-1}) in Fig. 1. This shows blueing on the long wave-length side, and reddening on the short wave-length side, of a minimum situated at $1/\lambda = 2{\cdot}3\,\mu^{-1}$ (λ 4350).

It must be remembered that this curve represents a combination of what are actually different extinction curves, since they are obtained from separate "blue" and normal solar spectrograms, taken at different times. Hence there is the possibility of varying conditions, e. g. a variation of the thickness of the absorbing layer in the case of the "blue" Sun, or a variation of atmospheric conditions in either case. The essential nature of the curve will, however, remain unchanged.

TABLE II

Observed extinction values of scattering layer

λ (A.)	$\frac{1}{\lambda}$ (μ^{-1})	$\frac{5}{2}\log_{10}\left(c_1\frac{I_N}{I_B}\right)$			$\frac{5}{2}\log_{10}\left(c_2\frac{I_N}{I_B}\right)$	α	$\frac{5}{2}\log_{10}\frac{I_N}{I_B}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
6338	1·578				0·03		10·92
6202	1·612				$\bar{1}$·96		10·85
6042	1·655				$\bar{1}$·88		10·77
5967	1·676	1·07			$\bar{1}$·87	−0·01	10·76
5803	1·723	0·93			$\bar{1}$·77	−0·05	10·64
5642	1·772	0·82			$\bar{1}$·68	−0·07	10·54
5535	1·807	0·74			$\bar{1}$·67	−0·14	10·50
5501	1·818	0·73			$\bar{1}$·66	−0·14	10·48
5382	1·858	0·73			$\bar{1}$·52	0·00	10·42
5246	1·906	0·66			$\bar{1}$·36	+0·09	10·30
5070	1·972	0·53			$\bar{1}$·28	+0·04	10·19
4900	2·041	0·46			$\bar{1}$·20	+0·05	10·12
4835	2·068	0·43			$\bar{1}$·16	+0·06	10·09
4744	2·108	0·41			$\bar{1}$·15	+0·05	10·07
4627	2·161	0·38			$\bar{1}$·12	+0·05	10·04
4550	2·198	0·35			$\bar{1}$·11	+0·03	10·02
4428	2·258	0·33			$\bar{1}$·09	+0·03	10·00
4370	2·289	0·32			$\bar{1}$·09	+0·02	9·99
4332	2·308	0·32			$\bar{1}$·09	+0·02	9·99
4206	2·378	0·33					10·00
4110	2·433	0·40	0·48				10·07
4050	2·469	0·42	0·52				10·09
4028	2·483		0·54				10·12
3981	2·512		0·61				10·19
3955	2·529		0·67				10·25
3951	2·531		0·68				10·26
3916	2·554		0·73	0·83			10·31
3893	2·569			0·97			10·45
3865	2·587			1·09			10·57
3845	2·601			1·16			10·64
3811	2·624			1·29			10·77

The points obtained from the extinction values given in columns 3 and 6 are denoted by full circles. These values can be combined, since they have been shown to have a reasonable correspondence by the values of α in column 7. The points obtained from the extinction values given in columns 4 and 5 are denoted by open circles and crosses respectively. These must necessarily be given less weight than the full circles. They cover a spectral region that was never meant to be utilized in the instrument used. The instrumental absorption is becoming heavy, as can be seen from the microphotometer tracings in Plate 8,

and consequently the gradient of photometric density is large; the spectrum is no longer in sharp focus and is in close proximity to the edge of the plate holder; the spectrum is over-exposed in the visible and hence the errors due to scattered light are magnified. Further, there is no check on the correspondence of these points with the full circles, and atmospheric variations are known to be violent in this region. The value of these two series of points lies, however, in the circumstance that each series indicates that the extinction is increasing with decreasing wave-length in the violet and near ultra-violet, thereby confirming the existence of the extinction minimum at $1/\lambda = 2{\cdot}3\,\mu^{-1}$.

The full curve in Fig. 1 represents a theoretically determined curve, the computation of which is explained later in the text.

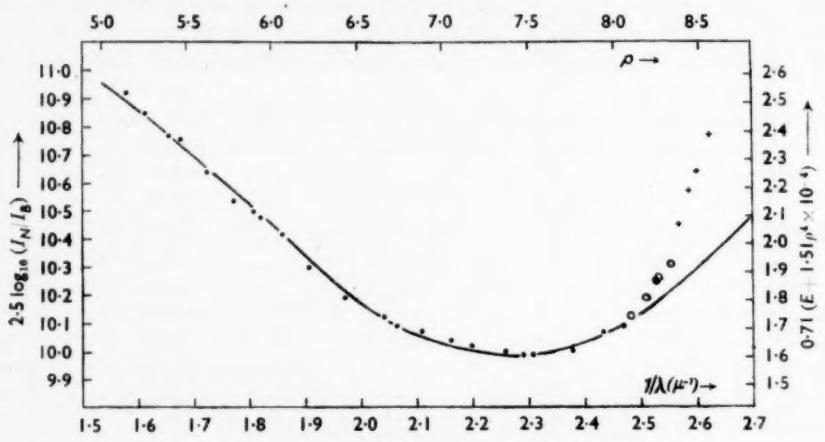


FIG. 1.—*Plot of observed extinction values against* $1/\lambda$.
Full curve represents theoretically determined extinction plotted against ρ.

2. *Resumé of theory.*—The problem of light extinction by a system of particles of known dimensions and electric and magnetic properties can in principle be solved on the basis of Maxwell's electromagnetic light theory. Due to the great labour involved, however, the computation of extinction curves has mostly been confined to the following case—scattering of a plane wave by a single isotropic sphere, the complete analytical solution of which was given by Mie (1). To apply these results to a cloud of isotropic spheres we have the added conditions that the source and observer must be at a large distance from the cloud, and the particles at a low enough density not to cause any interference with each other. If there is a dispersion in size and constitution of the individual isotropic spheres, the resulting extinction curve may be found by taking the mean of the various extinction curves for every type of particle and weighting with respect to population.

The results depend upon the two values r, the radius of the sphere and m, its complex refractive index relative to the surrounding medium.

$$m = \sqrt{(\epsilon - i\sigma/\omega)},$$

where ϵ is the dielectric constant and σ the conductivity of the material, i the imaginary unit and $\omega = 2\pi c/\lambda$, c being the velocity of, and λ the wave-length of the incident light in the surrounding medium. The magnetic permeability is assumed

to be unity. The extinction E is defined as the ratio of scattering to geometrical cross-section i. e. for a single spherical particle, E is the ratio of the scattered light to the light incident on the sphere. E is usually determined in terms of m and of the parameter $x = 2\pi r/\lambda$.

The computations are divided into three groups (a) m real; (b) $m \to \infty$; (c) m complex. Group (a) has received most attention and three curves are reproduced in Fig. 2. E is plotted against the parameter $\rho = 2x(m-1)$, which

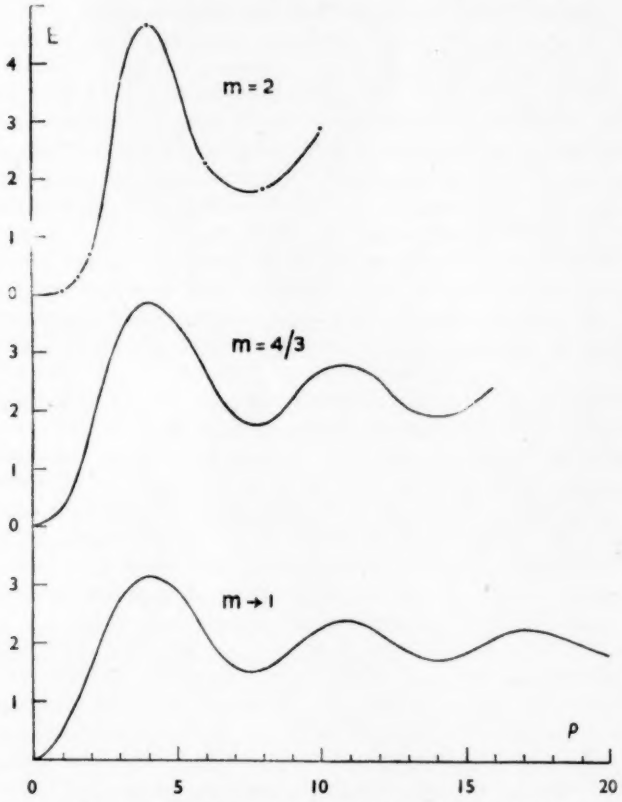


FIG. 2.—*Theoretical extinction curves for values of m real.*
E plotted against $\rho = 4\pi r(m-1)/\lambda$.
$m \to 1$: *Analytical formula by Van de Hulst.*
$m = \frac{4}{3}$: *Smoothed curve of computations by Houghton and Chalker.*
$m = 2$: *Computations by Greenstein.*

was shown by Van de Hulst (2) to be of more basic significance than x. It represents the phase shift suffered by a ray of light traversing the sphere along a diameter. Using this parameter as abscissa, the extinction curves for various m in the neighbourhood of unity are very similar. This is shown by Fig. 2 which is based on Fig. 21 of Part I of Van de Hulst's papers. The curve for $m \to 1$ is obtained from the formula

$$E = 2 - 4 \sin\rho/\rho + 4(1 - \cos\rho)/\rho^2$$

derived by Van de Hulst. This formula is for the actual limiting case $m \to 1$. For values of m near unity the formula gives a good representation of E except near the origin for those values of ρ small compared to $(m-1)$. For small values of x, Van de Hulst's formula gives $E = 2(m-1)^2x^2$, whereas the actual expression for E is of course the Rayleigh formula $E = 32(m-1)^2x^4/27$.

The curve for $m = \frac{4}{3}$ is the smoothed curve obtained from the extensive computations of Houghton and Chalker (**3**). The unsmoothed curve shows numerous minor fluctuations in the vicinity of the major maxima and minima, and it would require only a small dispersion in r to smooth these out. The computed points for the third curve ($m = 2$) are due to Greenstein (**4**). These are only sufficient, however, to give a smoothed version of the curve which preserves the major features. The more extensive computations of La Mer show minor fluctuations. The actual calculations of La Mer were not available to the author, but Van de Hulst plots them graphically in Fig. 3 of Part 2 of his papers. The curves start at $E = 0$ for $\rho = 0$, rise to a first maximum at $\rho \simeq 4$, reach a first minimum at $\rho \simeq 7{\cdot}5$ and then go through a series of diminishing undulations, which asymptotically approach the value $E = 2$.

The region for $m > 2$ is almost devoid of practical interest and hence few investigations have been carried out. The form of the extinction curve is therefore uncertain. As m increases above 2, the maxima depicted in Fig. 2 will be crowded towards the origin when x (not ρ) is used as abscissa. Moreover, Van de Hulst (**2**) has shown that an additional series of "resonance" maxima will develop for large m and small x. These approach the positions $x = k\pi/m$ (where k is an integer) and the values $E = \frac{1}{2}$, but the interrelation of these with the "interference" maxima, referred to above, is obscure. Van de Hulst, in his examination of the extinction curves obtained for a number of values of $m < 2$ from the numerous calculations of La Mer, has pointed out the small "bump" in the initial upward slope of the curves. In the case of $m = 2$, this is a definite subsidiary maximum. He shows that these "bumps" correspond to the position $x = \pi/m$ and suggests that they are related to the resonance maxima. It does indeed seem possible that they represent the development of the first resonance peak.

For the limiting case of group (b) $m \to \infty$, computations have been made by Götz (**5**) and Greenstein (**4**). This is the case of totally reflecting spheres. The curve based on the combined calculations of Götz and Greenstein is of a similar nature to those for m real and less than 2. It shows a rise to a first maximum and then undulates as it asymptotically approaches $E = 2$. The undulations are very much reduced, however, the first maximum having a value $E = 2{\cdot}29$ at $x = 1{\cdot}2$ and the first minimum having $E = 2{\cdot}12$ at $x = 1{\cdot}6$. Again, the relation of this region with that for m real and large (but finite) is obscure. The resonance maxima at $x = k\pi/m$ for large m and small x must still exist, but as m increases they all run together at $x = 0$ and disappear in the limiting case $m \to \infty$. This is a case of non-uniform convergence, the curves for m large converging non-uniformly to the limiting case $m \to \infty$.

For group (c) m complex, Van de Hulst has investigated analytically the case $m \to 1$ along the same lines as his investigations for dielectric spheres. He shows that the extinction is similar to the dielectric case on a ρ basis, but only a small conductivity is required to smooth out the undulations entirely. As the conductivity increases further, the first maximum also disappears and the curve rises asymptotically from the origin to the value $E = 2$.

3. *Comparison of observational material with theory.*—Returning to the present problem, it is reasonable to place the "blue" Sun extinction in the class m real. An extinction minimum can be produced by conducting particles, if the conductivity is small (as was indicated above), but the minimum will be much more shallow than for dielectric particles. It is therefore proposed that the observed extinction curve can be represented by curves of the nature depicted in Fig. 2. For the superposition of the extinction curves of the individual dielectric particles in the absorbing layer to produce a minimum, we must have small dispersion in the quantity ρ, i.e. small dispersion of $r(m-1)$, otherwise the various maxima and minima of the individual curves would overlap and the undulations be smoothed out. The only physical reason for $\delta\{r(m-1)\}$ being small is of course that δr and δm are small, i.e. the dielectric particles are uniform in size and constitution. An upper limit of the dispersion, assuming a uniform distribution, is given by the variation required to superimpose the first maximum and first minimum, smaller dispersions being required to superimpose any other consecutive maxima and minima. The respective values of ρ for the first maximum and minimum are 4·1 and 7·6 and hence the upper limit of dispersion is given by

$$\Delta\rho/\rho_{\text{mean}} = 0{\cdot}6.$$

This gives the upper limit as 60 per cent for $r(m-1)$, or assuming m to be constant, 60 per cent for r. The actual dispersion is probably much less, the minimum becoming very shallow as the dispersion approaches 60 per cent.

The observed minimum must represent the first minimum of the theoretical curve, since any other minimum would require the preceding maximum to be within the observed range. But a close comparison of the observed and theoretical curves did not show agreement. The slope of the observed curve on the short wave-length side of the minimum is steeper than would be expected from theory. Such an effect would be caused if, in addition to the uniform dielectric particle extinction, there was an overall reddening. Such an effect is to be expected in the comparison of the atmosphere at different times, the variation of normal atmospheric reddening having been shown (**6**) to be considerable.

There is also the possibility of instrumental colour effects. The effect of using different optical trains was examined in the laboratory. Since by far the greatest thickness of glass that the light passes through in the spectrograph is due to the two prisms, the effect of these alone was investigated. Light was passed through varying thicknesses of one of these prisms and the distribution of intensity throughout the spectrum measured by means of an Evans blocking layer photo-cell and galvanometer. From these measurements the absorption coefficient of the glass was determined at a number of wave-lengths. The absorption of a single pencil passing through a mean thickness of prism (the normal Sun) was then determined, and the absorption of a beam using the full aperture ("blue" Sun) computed. A comparison of the two showed agreement up to $1/\lambda = 2{\cdot}3\,\mu^{-1}$, beyond which the normal Sun is reddened with respect to the "blue" Sun. Hence the correction for this effect will increase the discrepancy in the observed curve. The correction a to be applied to the slope $dy/d(1/\lambda)$ of the observed curve, is given in the following table:

$1/\lambda(\mu^{-1})$	1·6–2·3	2·4	2·5	2·6
a	0·0	+0·3	+0·6	+0·9

The normal solar spectrograms were taken at later dates, at intervals of 30 and 50 days after the "blue" Sun. There is, therefore, the possibility of a change in colour of the instrument. This would be caused mainly by a deterioration of the silver coat on the primary mirror (the Cassegrain mirror of the Edinburgh reflector is aluminium coated), which has been shown by Greaves, Davidson and Martin (**6**) to redden with time. These workers, in an exhaustive survey of such effects, give a reddening of $+0{\cdot}2$ in terms of $dy/d(1/\lambda)$ for 30 days, this figure applying to a wave-length range 4400 A. to 6500 A. and to an optical train involving four silvered surfaces. Again, this will aggravate the distortion of the observed curve, which would require an instrumental blueing for explanation.

The possibility of vignetting can be ruled out. All the rays over the wave-length range examined enter the camera lens.

The instrumental reddening will cause an apparent blueing of the uniform dielectric particle extinction, and hence the atmospheric reddening of the "blue" Sun relative to the normal Sun must have been large enough to overcome this. Assuming an effective overall reddening, and that this can be represented by a $(1/\lambda)^n$ law, we have, in general,

$$y = 2{\cdot}5 \log_{10}(I_N/I_B) = 2{\cdot}5\,(\log_{10} e)\pi\Sigma(r^2NE) + p(1/\lambda)^n,$$

where N is the total number of scattering particles, of radius r, in a beam of unit cross-section and E is the corresponding extinction.

With no dispersion in r and since $\rho = 4\pi r(m-1)/\lambda$, the above equation becomes

$$y = k(E + q\rho^n),$$

where $\rho = h/\lambda$ and k, q and h are constants. It is now required to choose values of k, q, h and n, so as to fit the observed curve. The values of E are taken from the smoothed curve drawn through the computations of Houghton and Chalker for $m = \frac{4}{3}$. This curve is given in Fig. 2. The actual value of m for the scattering particles is unknown, but the similarity of the extinction curves for $m \to 1$ to $m = 2$ has been pointed out by Van de Hulst and is illustrated in Fig. 2. Accordingly, the curve for $m = \frac{4}{3}$ can be taken as roughly representing the extinction curve for the unknown m, when plotted on a ρ basis. In particular, the value of ρ, corresponding to the first minimum, varies little for different values of m, and hence the value of the constant h, obtained from fitting the observations to theory for $m = \frac{4}{3}$, should be closely correct (see below).

Values of y were computed and, by trial and error, the best fit was obtained firstly, for all the observations, and secondly, neglecting the observations beyond $1/\lambda = 2{\cdot}47\,\mu^{-1}$, which are of doubtful accuracy (see Section 1 above). The computations were confined to $n = 1$ and $n = 4$. No reasonable fit was possible in the first case (all observations), but in the second case a good fit was obtained (apart from a difference of zero on which comment will be made below) for $k = 0{\cdot}71$, $q = 1{\cdot}51 \times 10^{-4}$, $h = 3{\cdot}26$ and $n = 4$. This theoretical curve

$$y = 0{\cdot}71(E + 1{\cdot}51\rho^4 \times 10^{-4})$$

is plotted in Fig. 1 as the full curve.

The real overall reddening is easily explainable by normal atmospheric variation. The departure of the doubtful points from the theoretical curve seems great, even allowing for the experimental difficulties, but extra atmospheric reddening (the points are obtained from spectra taken at different times) at this altitude (19°), and in this spectral region, can become quite violent.

The value of ρ for the first minimum of the smoothed version of the calculations by Houghton and Chalker is found to be 7·6. The minimum of the reddened theoretical curve, obtained to fit theory and observation, is shifted to $\rho = 7{\cdot}5$, a shift of less than two per cent. Having regard to the general similarity of the curves for different m in the range $m \to 1$ to $m = 2$, it is reasonable to suppose that a similar shift (of the order of two per cent) would be found by using other values of m. Now the value $h = 3{\cdot}26$, quoted above, was obtained from the consideration that the value of $\rho(7{\cdot}5)$, for the first minimum of $E + q\rho^n$, corresponds to $1/\lambda = 2{\cdot}3\,\mu^{-1}$ for the minimum of the observed curve. It would therefore appear that the value $h = 3{\cdot}26$ is correct to a few per cent. But since $\rho = 4\pi r(m-1)/\lambda$, then

$$h = 2\pi d(m-1),$$

where d is the diameter of the scattering particles. Substituting for h gives the relation between the diameter and refractive index of the scattering particles to be

$$d(m-1) = 0{\cdot}519 \text{ microns}.$$

Values of d for various m are given in the following table:

m	1·10	1·25	1·33	1·50	1·75	2·00	3·00
$d(\mu)$	5·19	2·08	1·57	1·04	0·69	0·52	0·26

The value of k obtained above gives

$$k = 2{\cdot}5\,(\log_{10} e)\pi r^2 N_0 = 0{\cdot}71.$$

It is pointed out later that m is probably in the vicinity of 1·5 and therefore r is probably of the order of $0{\cdot}5\,\mu$. Using this, the order of value of the total number of scattering particles N_0 in a beam of unit cross-section, is determined as

$$N_0 \simeq 8 \times 10^7 \text{ cm.}^{-2}.$$

This value is for an altitude 19°. Hence, if n_0 is the average density of the scattering particles and S the thickness of the layer,

$$N_0 = n_0 S \operatorname{cosec} 19^\circ.$$

The thickness of the layer was determined to be 12,000 feet by an aircraft sent from Leuchars. This gives the order of density n_0 of the scattering particles to be

$$n_0 \simeq 70 \text{ cm.}^{-3}.$$

It will be seen from Fig. 1 that there is a considerable difference of zero, amounting to 8·4 magnitudes (a light intensity ratio of 2300/1), between the observed y and the theoretical $y = 0{\cdot}71\ (E + 1{\cdot}51\rho^4 \times 10^{-4})$. This indicates a heavy non-selective obscuration in addition to the selective extinction. Such a neutral obscuration, over the observed wave-length range, could be caused by particles of large size (i.e. large ρ) or by heavily absorbing (i.e. conducting) particles of smaller size (see Section 2). Part or possibly all of the effect may have been caused by cloud (water droplets with large ρ) at the time of the "blue" Sun for, although the "blue" solar spectrograms were obtained through a gap in low-lying cloud, cloud at a higher level may have produced obscuration. On the other hand, it is possible that part of the non-selective obscuration may be localized in the layer itself. The experimental evidence in favour of this interesting possibility is weak. It rests solely on the large value of obscuration (8·4 magnitudes) to be explained and the visual observation of the author that, at the time of observation, the "blue" Sun seemed clear of cloud.

4. *General discussion.*—To summarize the conclusions briefly, it is found that the layer causing the "blue" Sun has an extinction minimum, thus affording an experimental verification of an outstanding feature predicted by scattering theory. The results are shown to agree with theory on the assumption of an overall reddening. The scattering particles producing the minimum are dielectric in nature and remarkably homogeneous in size and constitution. A relation between the diameter and refractive index of the particles is deduced, and an order of their density is given. The possibility is also indicated that, in conjunction with the uniform dielectric particles, there is another system of particles in the layer producing a neutral extinction.

The source of the layer has been determined by Bull (**7**) as the extensive forest fires burning in Alberta on 1950 September 23. He states that the smoke from these fires reached the large cities of eastern Canada on September 24, and was thick enough there to blot out the Sun. The upper winds over the Atlantic were then favourable for bringing the smoke to Europe. A trajectory starting from the Toronto area on September 24, calculated from synoptic charts and including the upper wind observations of the Ocean Weather ships, would bring the smoke to Scotland on September 26. The smoke then travelled over to the Continent where the effects were seen until September 30.

Over North America the effect of the smoke was quite different owing to its great density. The fires occurred in muskeg country and are characterized by tremendous volumes of smoke. Elsley (**8**) gives an account of the visual effects and shows from wind observations that the smoke came from the fires. He states that at times there was complete darkness and the Sun, when visible, was purple or blue. He describes the colours of the sky (i. e. the scattered light) as shades of orange, pink and yellow. He also states that aircraft flying through the smoke were found to be covered with an oily substance on landing. Unfortunately, however, no one seems to have taken some of the oil into the laboratory for analysis.

The smoke was also observed by Dr H. S. Hogg (**9**) from southern Ontario, 2000 miles from the fires, on September 24 and 25. She states that the Sun could be seen at periods on both days, and describes it as a "pale, bluish-mauve disc".

Over Edinburgh the smoke was observed on the afternoon of September 26, when the Sun could be seen quite clearly through the haze and was a deep indigo blue. The extent of the layer was determined by an aircraft sent from Leuchars, which localized the smoke between 31,000 and 43,000 feet.

It seems, in the light of Elsley's (**8**) information, that the scattering particles are globules of oil, caused by the distillation of wood. Their uniform character is yet to be explained, and may have been established by combustion in the fire itself or by an atmospheric filtering process favouring particles of one size. As regards size, most organic oils have a refractive index near $m = 1{\cdot}5$ which would correspond to a diameter of the order of one micron. If, as has been indicated in Section 3, there is another system of particles in the scattering layer, producing a non-selective obscuration, their most likely constitution is carbon. This material is highly absorbent in the visible spectral range, and hence the extinction of such particles can be represented by a curve which rises asymptotically from the origin to the value $E = 2$, the extinction becoming practically neutral in the region of $\rho \simeq 4$ (see Section 2). The interesting possibility therefore arises that they can be of the same size as the uniform dielectric particles.

The occurrence of this "blue" Sun is not unique. There have been other cases, usually connected with dust storms in a desert, or a volcanic eruption. The most famous case is the eruption of Krakatoa in 1883, the effects of which were observed all over the globe.

The author would like to express his thanks to Dr E. A. Baker and Dr M. A. Ellison for their advice and encouragement, and also his gratitude to Professor W. M. H. Greaves for the very many suggestions and criticisms he made throughout the whole of this investigation.

Royal Observatory,
Edinburgh:
1951 *May* 28.

References

(1) G. Mie, *Ann. d. Physik*, **25**, 377, 1908.
(2) H. C. van de Hulst, *Recherches Astronomiques de l'Observatoire d'Utrecht*, XI, parts 1 and 2, 1946, 1949.
(3) H. G. Houghton and W. R. Chalker, *J. Opt. Soc. America*, **39**, 955, 1949.
(4) J. L. Greenstein, *Harvard Circular*, No. 422, 1937.
(5) F. W. P. Götz, *Astr. Nachr.*, **255**, 63, 1935.
(6) W. M. H. Greaves, C. Davidson and E. Martin, *M.N.*, **94**, 488, 1934.
(7) G. A. Bull, *Met. Mag.*, **80**, 1, 1951.
(8) E. M. Elsley, *Weather*, **6**, 22, 1951.
(9) H. S. Hogg, *J.R.A.S. Canada*, **44**, 241, 1950.

DELAYS IN FLUORESCENCE APPLICABLE TO THE BLUE-RED BINARIES T COR B, Z AND R, R AQAR, AX PERS, α SCOR AND RW HYDR

Martin Johnson

(Revised version received 1951 January 5)*

Summary

In enquiring where the lines of high excitation originate in a blue-red binary system, and in seeking the mechanism of variability in those lines, it is desirable to compare times of delay in rise of fluorescence with the time of fluctuation of the stellar stimulus. Equations of Grotrian, Zanstra, and Cillié are therefore employed to correlate rates of growth of fluorescence with certain of the circumstellar conditions, applied to spherical models emitting the Balmer lines and 4686 A., and also the *O* III and *N* III fluorescence excited by 304 A. of *He* II. Data as to observed rates of change, from Swings and Struve, Merrill, Payne-Gaposchkin, McLaughlin, and Sanford, are compared with the suggestions offered by the models. In T Cor B, Z Andr, R Aqar, AX Pers, the comparisons display the contrast between fast-changing inner nebulosity and slow distant relics of previous outbursts, and facilitate the provisional location of the bright lines in each system. On the other hand, the more steady α Scor and RW Hydr illustrate, from the distance of the M component and its low amplitude of pulsation, the connection suggested by Swings and Struve between binary properties and fluctuations.

1. *Introduction.*—A number of peculiar stars which show molecular spectra, and yet high excitation and also nebular lines, are commonly explained as the binary association of a cool and perhaps pulsating M giant with a hot B star; the pair seems liable to nova tendencies and capable of stimulating nebular fluorescence. The problem arises why this behaviour should depend on the binary character, speculation by various authors having contrasted these cases with somewhat similar blue stars which lack the red companion but thereby fail to produce either the circumstellar nebulosity or the occasional explosiveness.† Understanding of any such dependence on the red companion awaits the solution of two problems particularly:

(*a*) the location of the bright emission between or around the pair of stars;

(*b*) the cause of *variability* in the emission, which remains undecided even after Bowen, Struve and others have elucidated much of the fluorescent process of bright lines in its more nearly steady features.

In the present treatment we begin by classifying some possible sites for the bright emission and some causes for variability which would be appropriate

* Received in original form 1950 October 31.

† I am indebted to Fellows for emphasizing that these stars cannot be guaranteed as binary until periodic velocities of a blue star can be extracted from the complex variety of displacements in the circumstellar strata. The argument of this paper as to shell distances would apply also to any hypothetical single star which could provide the combination spectrum; but I write in terms of the usual binary hypothesis until a definite alternative is propounded, not even adopting the caution of authors who have placed " companion star " in inverted commas. It does not seem that absence of orbital shift is contrary evidence, when these great distances of circumstellar materials are likely to preclude their sharing in the individual motions of the hidden but reasonably inferred blue companion.

in each site. As one means of discrimination between these possibilities, we then calculate the rate of growth of fluorescence in a gas located in any of these possible sites, when subjected to fluctuations in the source of its excitation. This rate of growth will be one factor in the time-development of an observed spectrum, a factor calculable *a priori* through selected models. It may become feasible to decide whether we find ourselves watching not the rate of change in the variable star but the later responses in its circumstellar gases. We apply our calculations, on principles due to Grotrian, Zanstra, and Cillié, to the large body of data now available concerning six peculiar stars likely to come under this blue-red category; by showing how much allowance is needed for this "time of adjustment" in models of various dimensions, the way may be cleared for assigning the observed rates of spectroscopic change to ascertainable mechanisms located in any particular part of such a binary system.

2. *Location of the bright lines of high excitation.*—In addition to the Balmer lines, these include *He* II, [*O* III], [*Fe* II], [*Fe* VII], and also the *O* III and *N* III which are excited locally by lines of *He* II through Bowen's mechanism (1). Possible sites for these emissions are as follows:

(i) in contemporary ejecta from the hotter star: such extrusion is understandable in terms of the hydrogen mechanism of single Be stars studied by Gerasimovič and others;

(ii) in the lower atmosphere of the M star: this would be possible if lines were sited among the emission spectra observed by Joy for *o* Ceti and treated theoretically by Scott;

(iii) in the outer envelope of the M star, where pulsation periodically varies the area of fluorescent target for irradiation by the B star;

(iv) in a mingling of (i) and (iii), especially if the pair is a close binary surrounded by the extruded ring or tail envisaged by Kuiper;

(v) in shells now detached and independent of the orbital distribution of the pair, if the B star is a post-nova; and

(vi) for stars of Gaposchkin's "nebular variable" class, in interstellar cloud not derived from either star.

3. *Sources of variability for the bright lines.*—Three kinds of agency will control any time-variation in the bright lines:

(I) any intrinsic variability in the B star initiating fluorescence in the above locations;

(II) geometrical factors altering the target area; and

(III) secondary physical modifications in the target.

Included under (I) are repetitive nova outbreaks, sporadic coronal and prominence activity on a large scale, slow secular changes in internal energy generation and stability, and long-period fluctuations such as were envisaged by Gerasimovič in "relaxation oscillations" when ejecta have become retarded after several years.

Under (II), even a constant B star can stimulate periodic fluorescence when eclipses of the B star by the M star impose successive shadow and re-illumination upon successive sectors of the targets in (iv), (v) and (vi), and also when the pulsating M star interposes in the path of the steady flux from B an expanding and contracting target, (iii) and (iv).

Among causes of variability under (III), there is the opaque molecular association or condensation which seems to develop among the ejecta of slow

novae; there is also the unpredictable drift and aggregation in interstellar clouds of non-uniform density if the lines are located in (vi).

4. *Orders of magnitude for rapidity of changes affecting the spectra.*—In assessing the contribution of these sources of variability to any observed rate of spectral change, we define as t_1 the time taken for the radiating B star to alter, t_2 the time for such source to become uncovered with respect to any portion of the target after eclipse, t_3 the time for the target to alter its area through pulsation of the M star, t_4 the time for physical modification in opacity.

In addition to these times for creation or exposure of source and for creation of fluorescent target, there is a further time defining our quantity τ, due to the fact that the visible consequences of the irradiation are not instantaneous; indeed, the mechanism by which a gas responds to a given new intensity of radiation may occupy a time negligible in the laboratory but weeks or months in circumstellar gases, as many writers have qualitatively suggested.

Our problem therefore becomes the investigation of the ratio

$$\tau/t_1, t_2, t_3, t_4$$

to decide whether any observed rate of change denotes growth of star or target, or "time of relaxation" delaying approach to equilibrium after both target and source have been established. It will be useful, before investigating τ for various models, to list (Table I) the rough orders of magnitude to be expected of the times in the various causative mechanisms, geometrical and physical. r is the likely distance, star to target, which we require later.

TABLE I

Orders of times for sources (I), (II), (III), and locations (i) . . . . (vi)

	t_1	t_2	t_3	t_4	r (cm.)
	I	II	II	III	
(i)	Few hours	...	...	e. g. day in ρ Cass	10^{11}–10^{12}
(iii)	...	...	Pulsation ½ period, e. g. 150 days	...	$1{\cdot}5 \times 10^{13}$
(iv)	...	Days in contact binary / Weeks in separated binary	" " "	...	$0{\cdot}5 \times 10^{14}$
(v)	...	Days or weeks	...	e. g. weeks in T Cor B / e. g. months in N Herc (1934)	10^{15}
(vi)	...	" " "	...	...	10^{17}

5. *Grotrian's theory.*—A method for evaluating τ was devised by Grotrian (**2**). He postulated the simplified model of a sphere of atomic hydrogen in its normal state, at whose centre a source of radiation is either newly established by rise in temperature of a star or uncovered by penetration of an opaque gas between star and outer nebulosity. The time for establishment of source was supposed small compared with time for growth of fluorescent response.

We first retrace some portions of his argument which need explanation, in order to see what remains valid when we apply the theory to another situation. We shall use the word "fluorescent" for all line emissions arising in regions not describable by the thermodynamic ideal of an enclosure in equilibrium, or even by the modifications thereof adequate to reversing layers: some writers have restricted the term to radiations excited by line sources, calling "recombination spectra" those which directly follow photo-ionization, but we can conveniently treat together instances of both kinds where line emission is in frequencies lower than that of the absorption which initiates the process.

We exchange Grotrian's notation for that familiar in experimental ionic recombination. If q is the number of ionizations per cm.^{3} per sec. by some agency, and α is a recombination coefficient whose dimensions can readily be found to be L^3T^{-1} if n is the concentration of ions per cm.^{3},

$$dn/dt = q - \alpha n^2. \tag{1}$$

It is usually convenient to define

$$k^2 = q/\alpha,$$

so that

$$\left.\begin{aligned} \int_0^n \frac{dn}{k^2 - n^2} &= \int_0^t \alpha\, dt, \\ \left[\frac{1}{2k} - \log\frac{k+n}{k-n}\right]_0^n &= \alpha t, \\ \frac{k+n}{k-n} &= e^{2k\alpha t}, \\ n &= k\frac{e^{2k\alpha t} - 1}{e^{2k\alpha t} + 1}, \\ n_\infty &= k = \sqrt{(q/\alpha)}. \end{aligned}\right\} \tag{2}$$

Hence the quantity τ, referred to above, can be defined

$$\tau = \frac{1}{2k\alpha} = \frac{1}{2\sqrt{\alpha q}}. \tag{3}$$

Swings and Struve (**3**), and later the present author, followed Grotrian in using this τ in describing the growth of luminescence in a nebulosity during development of recombination after absorption of ionizing radiation from a central star.

Since observable luminosity of this kind depends on n^2 in cases where electronic and ionic concentrations are comparable, Grotrian plotted the recovery in luminosity of N Herc (1934) from the equivalent formula

$$I_t = I_\infty \left[\frac{1 - e^{-2\sqrt{\alpha q}t}}{1 + e^{-2\sqrt{\alpha q}t}}\right]^2. \tag{4}$$

Grotrian's dependence upon any considerations beyond common theory is usefully separable into (*a*) a transformation of the dimensions of α to take account of the extension of circumstellar gas, (*b*) a corresponding change in the dimensions of q. The validity of each of these might fail for different models, so we shall derive quantities A and Q of the required dimensions to be substituted for α and q in the general recombination theory of equations (3), (4).

(*a*) For his applications, Grotrian needed to calculate not the n ions per cm.^{3} but N the total number in the circumstellar gas which he had represented as a model sphere of radius r. Accordingly the recombination coefficient α, which we used above in the manner defined in treatises on electronics, has to be replaced by

$$A = \alpha/\tfrac{4}{3}\pi r^3. \tag{5}$$

This quantity can play the same part as was assigned to α in equations (3) and (4), with dimension T^{-1} replacing the original L^3T^{-1}. Using Cillié's theory (**4**) of hydrogen recombination, Grotrian proceeds by a stage equivalent to writing

$$A = C\frac{M(m, T)}{\frac{4}{3}\pi r^3}, \tag{6}$$

$$C = \frac{2^9\pi^5}{(6\pi)^{3/2}}\frac{\epsilon^{10}}{\mu^2 c^3 h^3}\left(\frac{\mu}{k}\right)^{3/2} = 3{\cdot}24 \times 10^{-6}. \tag{7}$$

The order of magnitude of r in relevant cases we listed in Table I.

(*b*) The ionizations per cm.^{3}, q, is similarly replaced by Q, the total number of ionizations in the model sphere, for which Grotrian uses Zanstra's tabulation (**5**) of the number of quanta, whose frequency exceeds the minimum for ionization, emitted per sec. from a spherical stellar black-body radiator of radius R, with no absorption lines within that particular continuum. The identification of number of ionizations in the gas sphere with number of quanta emitted by the star we discuss below.

$$Q = \frac{8\pi^2 R^2 k^3}{c^2 h^3} T^3 \int \frac{x^2}{e^x - 1} dx \quad (x = h\nu/kT). \tag{8}$$

In our previous application (**6**) to several stars, one of which, Z Andr, links with the present different problem, values of τ as defined by (3) were computed: for the present purpose it is more useful to compute the course of I_t/I_∞ from an equation of type (4) but using A and Q from (6) and (8) to replace α and q of the general theory, as in Grotrian's treatment of N Herc (1934). We thus discover at what times recognizable fractions of final fluorescence are attainable in particular models. In the first of our present applications we assume the B star irradiating its surrounding gases to have radius 10^{11} cm., and we plot (Fig. 1) I_t/I_∞ against $\log t$ for each target radius r expected in the different hypothetical locations of bright lines in Table I. The temperature for the radiator is an early B star value of 20,000 deg. and this is also taken for the electron temperature in the capture zone of the circumstellar gas. We use in (6) at this stage Cillié's tabulated functions for $M(m, T)$, which apply to Balmer lines; we give the derivation below in (10) when we need to recalculate its value for another gas. We use Zanstra's tabulation for Q, following Grotrian. Comparison with t_1, t_2, t_3, t_4 of Table I can then indicate, for any selected effect of star on surrounding gas, where the rate of growth of response to irradiation becomes large or small compared with the time for creation or exposure of source or target.

6. *Limitations of Grotrian's model.*—The following considerations were not discussed by Grotrian, and must be borne in mind before we proceed to fresh values of $M(m, T)$ and $h\nu$ in (6) and (8).

(i) The original model is of an atomic hydrogen sphere surrounding the erupting star, the boundary of the sphere being defined by radial extension sufficient for total absorption; this definition allowed Grotrian to equate the number of ionizations to number of available quanta. Accordingly, although concentrations of material do not appear explicitly in his equations, any replacement of one sphere by another of greater density diminishes this volume and vice versa. For spheres whose actual boundary of their material fell short of the optical depth for tota' absorption, the calculated τ would be an overestimate: on the other hand, the actual non-uniformity in concentration of any one gas means that the simple exponentials (Figs. 1, 2, 3) are probably too gradual for the inner portions of the circumstellar shell and too steep for the outer. If different gases exhibit differing density gradients, Grotrian's volume of "total" absorption will have a different radius for each gas, the fluorescence emerging from different depths; but it is probable that such observed "stratification" in bright-line spectra is more commonly due to monochromatic attenuation, as mentioned below in (iii), than to any separating of gases under differential radiation pressure, the smoothing of which by viscous drag has been discussed in a previous paper (**6**) based on McCrea's equations.

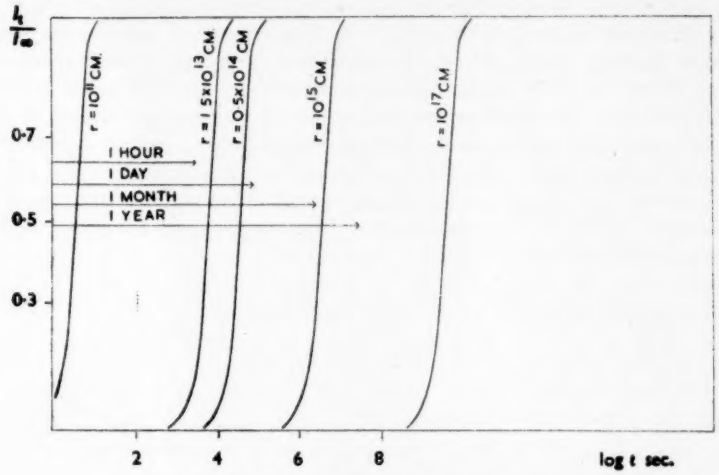


FIG. 1.—*Time of growth of Balmer lines at different distances.*
Capture temperature : 20,000 deg.
Radiation temperature : 20,000 deg.

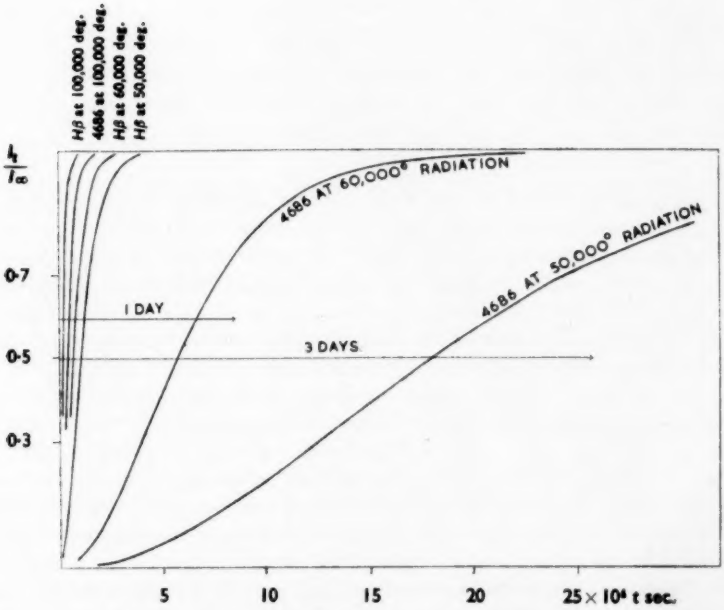


FIG. 2.—*Time of growth of Hβ and* 4686 A. *at* 10^{14} *cm. and* 10,000 *deg. capture temperature.*

(ii) Distortion of the model curves will also result from any divergence of radiation intensity in any wave-band from that assumed in (8). For the very important factor of general "dilution" by inverse-square attenuation, there is equivalent provision in (5), where the expression for recombination which affects the time of growth of fluorescence is dominated by the r bounding the volume of total absorption. The dilution, often mentioned in the American work which we quote on circumstellar gases, becomes by definition infinite at Grotrian's boundary r. The luminosity is a function of both the ionization and the subsequent recombination, so that diminution both of radiation and of gas density with distance makes any calculation of I_t/I_∞ a rough approximation, failing as the boundary is approached.

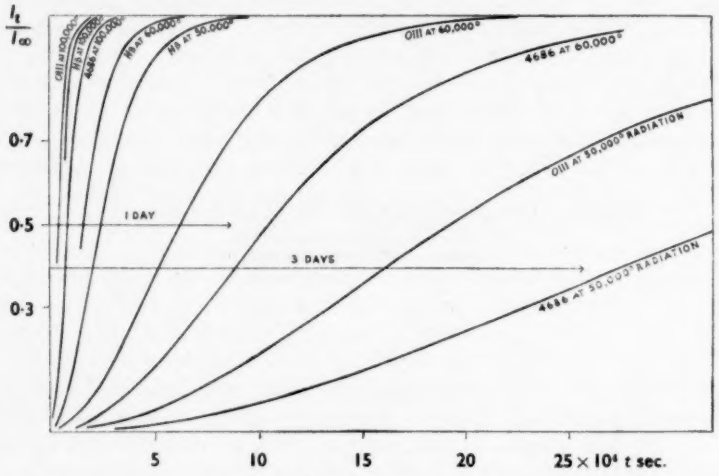


FIG. 3.—*Time of growth of Hβ*, 4686 A. *and O* III *at* 50,000 *deg. capture temperature.*

(iii) In addition to this general attenuation of all radiation with distance, and apart from non-uniformity in the gas mixture, monochromatic attenuation is a main source of the "stratification" which has been revealed by many observers: a material whose opacity is slight for any particular wave-band will present a larger radius for Grotrian's sphere, while high opacity in another frequency shortens the radius for some other material. The emission of different spectra would therefore mark out spheres of different radii even if the atomic concentrations were identical. In the opposite sense to the attenuation imposed by selective opacity, local intensification above the normal value can occur in certain frequencies: an important example which we shall use is the local excitation of *O* III and *N* III by recombination lines of *He* II according to Bowen's principle. A way to treat this will be to regard it as accompaniment of a fluorescence of *He* at 304 A. instead of at the visible 4686 A., and therefore of different calculable probability.

(iv) Other limitations to Grotrian's model include the assumption of equal ionic and electronic concentrations; but profitable comparison between different spectra may be made without removing further restrictions, so long as we recognize the limits which they set to claiming likeness between model and actual star.

7. *Delay in sphere model for He* II *and its local excitation of O* III *and N* III.—The above limitations introduce the important novelty that the causes of "stratification" will involve time-dependences of their own; when the central source of radiation fluctuates, the response in fluorescence may exhibit a succession of spectra from different radial distances, the intensity of each growing at a different rate with their different values of the quantities in (4). Optical strata will appear dispersed along a time sequence, even apart from the final stratification according to dilution of radiation and successive ionization potentials first studied by Bowen and extended by Beals to "shell" stars. These strata, we have suggested, indicate differences in Grotrian's volume of absorption for different gases. Such time sequences perhaps dominate the later or nebular stages of novae as well as the binaries which we consider in this paper, whereas at the beginning of a nova the observed succession of spectra may be determined mainly by changes in T and p of the erupting gases themselves.

We proceed to show the extent to which differences in A and Q in (6) and (8) are capable of imposing differences in rate of growth between different bright lines important in the blue-red binaries. In order to exhibit the effect of applying Grotrian's theory to other than Balmer lines, the differences in electron capture and in radiation intensity at the relevant frequency will be considered alone, divergence from a common value of r due to actual distribution of different gases being a problem not accessible to known data at present.

The following changes in the quantities are required.

(i) Electron capture theories, developed by Cillié from the method of Gaunt and ultimately of Kramers, will all involve functions of atomic number Z, hitherto taken as unity, so that for any atom generally

$$\alpha = \alpha_H Z^4 f(g, \dots), \tag{9}$$

where g denotes statistical weight. The only simple case is that of "hydrogen-like" ions, of which He^+ is relevant to these particular stars, and for this instance much can be inferred by including in the function f only the new values which have to be calculated in (6) for the number of captures on the mth level. Cillié's derivation of the dependence of this upon m and upon χ_m, the ionization potential from the mth energy level, is

$$M(m, T) = T^{-3/2} m^{-3} e^{\chi_m/kT} Ei(\chi_m/kT), \tag{10}$$

the last term being that tabulated under $-Ei(-x)$.

(ii) In (8), where $x = h\nu/kT$, transformation to He^+ implies a frequency limit corresponding to 228 A. in place of the 911 A. of hydrogen, but selection of m, the level at capture, depends (Table II) on the particular line observed or inferred from observation of its secondary consequences.

Hence, where we employed for Fig. 1 Cillié's own tabulation of $M(m, T)$ for the sum $m = 2 \dots 10$, a visible Balmer series, we shall now compare instead the appearance of (*a*) the single line $H\beta$, (*b*) 4686 A. of *He* II, (*c*) 304 A. of *He* II inferred from observation of the consequent excitation of *O* III and *N* III. Then the relevant values of the quantities are as follows, m_1 denoting the capture level, from which χ_m is the ionization potential, and $m_1 \to m_2$ the transition selected for giving the observed line of that element or of another element excited from the invisible line by Bowen's mechanism. Since the visible "line-excited fluorescence" in *O* III and *N* III covers only a few of the possible consequences of the lower level capture in He^+, it is obvious that our investigation of rate of attainment

of illumination cannot alone decide which line becomes ultimately the brightest, and we can compare only the *relative* delays in the emergence of the several spectra.

TABLE II

	m_1	m_2	χ_m
(*a*)	4	2	0·846 e.V.
(*b*)	4	3	3·38
(*c*)	2	1	13·54

We compute the function $M(m, T)$ for these values of m and χ_m, and we select the model of $r = 10^{14}$ cm. from the set of Fig. 1, as these extended nebular shells seem a common site for *He* II with *H*. Fig. 2 then shows the result of combining values of *A* at circumstellar electron-capture temperature of 10,000 deg. with photospheric temperatures of 50,000, 60,000, 100,000 deg.; these very high values are selected because *He* II characterizes the "hottest" eruptors. Fig. 3 shows the comparison between visible 4686 A. and the inaccessible lower level capture by He^+ which supplies the stimulus for observable *O* III and *N* III. Here the capture temperature is 50,000 deg. and the stellar temperature 50,000, 60,000 and 100,000 deg.

Values of the time abscissa are marked at intervals of an hour, a day, three days, a month and a year, and it will be recollected that the effect of differing sizes of Grotrian's model "totally absorbing sphere" is to distort any of these graphs laterally: a denser gas or a wave-band suffering less dilution would thus shift a curve towards briefer time taken for fluorescence to become strong, while density gradients along r will be further sources of distortion from the simple models.

Applications

8. ***T Cor B compared with Z Andr.***—Grotrian's "relaxation-time" was first applied by himself to the secondary brightening of N Herc (1934) during 200 days, then by Swings and Struve to the four weeks of recovery in the outer nebulosity of Z Andr in 1940, whose inner screen was similarly assigned a more rapid progression in our earlier paper. The present applications are suitably introduced by means of comparisons with Z Andr.

The outbreak of T Cor B in 1946, which McLaughlin (**7**) has shown to resemble in many ways its outbreak in 1866, exhibits the type of secondary rise in luminosity after dimming which has also characterized the slow novae of 1891, 1934 and 1942. It is possible that this apparent double rise is connected with Rosseland's and McLaughlin's mechanism of shock waves; but since it occurs in a minority of novae, another explanation applied to those cases by Chandrasekhar, Stratton, and ter Haar is worth exploring in T Cor B, whose light-curve is qualitatively similar. This explanation is that opaque clouds among the ejecta later become dissipated, causing the "secondary rise", which is thus really a recovery from transient dimming. Any such theory would be reinforced by what is known to occur in Z Andr, and would imply a mechanism of variability of the type (III) in our Section 3. In the light of this suggestion we proceed as follows.

Sanford (**8**) has compared Pettit's light-curve, during this "recovery", with the rise and fall in intensity of *H*, *He*, *He* II, *O* III, *N* III and [*O* III]; the minima of the recombination spectra and fluorescence occur around June 8,

whereas the recovery of total luminosity had begun about May 20, the peaks in bright lines being delayed by times varying over some weeks. If these delays are of Grotrian's type, our Fig. 1 indicates that the observed growth of hydrogen lines would correspond to a distance 10^{14}–10^{15} cm. This distance suggests the locations (iv) and (v) of our classification in Section 2, the late recovery of light denoting, as in Z Andr, a penetration of opaque gases between blue star and outer nebulosity. The outer may be partly drawn from the extended atmosphere of the M star, but in both Z Andr and T Cor B there was available material in the circumstellar remains of *recent* outbreaks (location (v) of Section 2): T Cor B differs from Z Andr (*a*) in its target for fluorescence provided by the very fast ejecta from February 10 in the same cycle, and (*b*) in its inner penetrable cloud not being identifiable as a small P Cygn shell such as Z Andr had extruded in 1940. A comparison with the quiet blue-red binary α Scor (Section 11 below) indicates how very widely separated a pair must be to maintain the M star's atmosphere uncontaminated.

The fact that *He* II, and the *O* III and *N* III related to it, rise before *H*, is possibly connected with our Section 6 and Figs. 2, 3. At the same temperature and same distance these spectra would develop more slowly than *H*, but in a more compact atmosphere of these gases they would develop more rapidly. This latter is likely in the case of T Cor B, in view of the extreme speed of *H* which must have scattered it to a great distance after its extrusion at 4000 km./sec. on February 10. In Z Andr in 1946 the line 4686 A. rose to equality with the Balmer spectrum during the fading after maximum light, which in our mechanism could be associated with the less extreme disparity between values of r for different gases, since hydrogen velocities at outburst were not so high as in T Cor B. McLaughlin also, like Sanford, records for T Cor B the overtaking of *H* by *He* in intensity some months earlier, during the initial motions of February; but this is not so readily ascribable to Grotrian's type of sequence as is the subsequent recovery, the *initial* rise of any nova being more nearly a function of T and p.

Correlation of the outburst with the M pulsation cycle of 217 days in T Cor B is not so clear: the M spectrum was naturally in February rendered invisible, and again weakened when fluorescence from the blue was recovered in June to August, but it seems uncertain whether the M radial velocities (with range as great as 48 km./sec.) are orbital or pulsatory, so their rates of change cannot yet be proved separable into our t_2 and t_3 (Section 4) as distinguished from Grotrian's τ.

9. *R Aqar compared with T Cor B and Z Andr.*—In R Aqar, Payne-Gaposchkin (**9**) considered that both blue and red components of the binary were periodic variables, the blue up to 5 mag. and the red of even greater amplitude. In the absence of any independent clue as to whether the lines of high excitation are actually located in the ejecta from the B star or in the atmosphere of the M star irradiated by B, the indication that the blue star is also periodic may mean that lines in material of M excited by B fluctuate as the target alters, without the source of radiation altering: the apparent "blue" periodicity may be an effect of steady radiation on a pulsating target which simulates a hot atmosphere by fluorescing when exposed to the blue star. A feature contrasting with Z Andr is that the amplitude of the M star is damped during the "active" times of the B star, the range falling to 2¼ mag., and the normal minimum is recovered sooner than the normal maximum, whereas in Z Andr the amplitude of M is least when B is faint. In the

present treatment this contrast is not anomalous, since a screening by transient opacity in the ejecta might affect the apparent visual luminosity of either or both stars of a pair, according to the constitution of the resultant shell and its distribution around one or both components.

Merrill (**10**) and Swings and Struve (**11**) record the intensity of nebular lines (including [*Fe* II] and 4959–5007 A.) as fluctuating widely with no obvious correlation with phase of M in R Aqar, and it has been commonly considered that the *He* II is here stellar and not in the external nebulosity. The blue star has been known to alter as rapidly as its great brightening of 1926 followed by a faintness in 1927, and again as slowly as the 1928–33 return to 1926 type. This type itself had changed from 1922–3, when in extreme faintness of the M star the whole system had simulated a planetary nebula and nucleus, such as NGC 6572, though at times there have been violet edges of outwardly moving lines nearly of P Cygn type, and Hubble discovered in the 1930s an outward motion persisting. At other times the richness in [*Fe* II] has been comparable with that of WY Gemi.

This lack of correlation between nebular fluctuation and phasing of M star suggests that some of the former is distant enough for Grotrian's lag to exceed the time t_3 of Table I and Section 4, a time which in this case arises from a period of 387 days. Reference to our Fig. 1 shows that times thus exceeding a year correspond to target-sizes very large, of order 10^{16} cm., for the hydrogen model. It is significant, however, that some lines seem less distant, since Swings and Struve show, from fluctuations in the ratio 4959/5007, that these can be occulted by the *TiO* atmosphere of the M star; this suggests the location (ii) of Section 2, as when Joy found *TiO* overlying Balmer emission in *o* Ceti, or at least that nebulosity is indeed located between the stellar components.

The discrepancy suggests the conception of *two* visible regions of nebulosity, one much smaller than the other and taking far less time to adjust itself to radiation: if this is correct, the situation becomes comparable to the inner P Cygn shell and outer "old" nebulosity of Z Andr, and also comparable to the similar view of T Cor B which our treatment has taken, where an inner veil suppressed the light-curve in 1946 March to May before penetration in June permitted illumination of an outer ejected nebula. But the case of R Aqar has the unique advantage that the two nebulosities yield photographic evidence not available in the other stars, notably in the hands of Hubble and Humason and Lampland (see, e.g., P. W. Merrill, *Spectra of Long-Period Variable Stars*, Chicago, 1940). The inner is a slight elongation of the stellar image, but the outer is a complex web lying at some distance outside the stellar region. The fact that the inner varies in place and brightness, while some features of the outer remain invariant, accords with the above estimated locations of the nebular lines, since the inner at the distance of an M star's dimension would appear from Fig. 1 to have a delay in days or weeks, but the outer may be at planetary nebular distance. If 10^{17} cm. is suggested for the latter, Grotrian's lag would amount to years in our graphs. Lampland's irregular density distribution in the outer nebula would cause overlapping in the several successive emergences of fluorescence from different materials, a fact also contributing to the lack of detectable variability in integrated light.

10. *AX Pers and CI Cygn as binaries of highest excitation.*—Similar evidence from observed time sequences, towards locating the source of bright lines, is

obtainable for these two, which are commonly rated as the highest in excitation among the blue-red binaries. The strength of [*Fe* VII] and probably [*Fe* X] led Swings and Struve (**12**) to postulate a stellar temperature of 150,000 deg., with a spectral maximum at 200 A. rendering the nucleus visually faint, but lines apparently coronal may represent local eruption. It is, however, certain that the great strength of *He* II relative to weak *He* I, recorded also by Merrill, places the blue component among the hottest known. The strength in [*Ne* V] and [*Fe* VII], resembling Pd type and N Pict (1925), makes AX Pers as likely a post-nova as any of these stars, but the sharpness of the lines shows that expansion is no longer fast. It is therefore a case favouring especially the location (v) of Section 2, rather than (i), i. e. bright lines in ejecta from a long-ago completed eruption, but with a nucleus retaining (or having now regained) radiating power in the extreme ultra-violet. This would not follow the Gaposchkins' suggestion in 1938, classifying CI Cygn as example of the "nebula possibly the outer envelope of the red star".

Swings and Struve suggested that the very hot nucleus of AX Pers may possess the [*Fe* VII] nebulosity close around itself, with only diluted radiation exciting the *Fe* II in the outer parts of the M star: the problem thereby aligns itself with that of Z Andr, R Aqar and T Cor B, in each of which two nebular regions of small and great distance respectively were distinguished. Here there may even be three. If in AX Pers the innermost is known by its response in [*Fe* VII] to undiluted excitation, the inner of T Cor B is only to be inferred from the time-lags which we have noted in a luminosity curve, if the latter's transient minimum denotes an opaque screen.

For some years CI Cygn and AX Pers resembled one another more than they resembled any other object, but after 1939 various changes in AX Pers have separated them, and some are relevant to the time-lags which we are discussing: these changes are recorded by Swings and Struve (**13**), Merrill (**14**) and Payne-Gaposchkin (**15**), and the following seems the reasonable classification suggested by our theory of time delays in fluorescence. (*a*) The lag in the nearby [*Fe* VII] would be so small in any material resembling our models that the loss in 1941 January regained in May and more fully in August must denote t_1 of Table I, not Grotrian's τ. This seems to be an instance where observed rate of change may denote the changing of source itself rather than delay in response, the luminosity keeping pace with the excitation ((I) of Section 3). (*b*) At the other extreme, the likeness to a planetary nebula even so small as IC 4997, requires adjustment times measurable in years on any of our figures, reference to which shows that delays of that order will occur when r rises towards 10^{17} cm. (*c*) Between these, changes observed in *H* and *He* II are more likely to represent the weeks or months of delay in response when the target dimensions are comparable with the binary orbit and the extent of the M star ((II) of Section 3). Associated with these changes are Payne-Gaposchkin's records that loss of 4686 A. at a low maximum in pulsation (i. e. presumably when M is veiled by ejecta from the hotter star) is followed by a rise towards equality with Balmer lines 400–430 days later. It is notable that Merrill's list of radial velocities shows an excess of negative velocity for *He* compared with *H*, not unexpected at this nuclear temperature, which must accelerate *He* beyond *H*. This may well provide a more distended and rarefied volume of *He* than of *H*, in contrast to the distribution in stars of lower excitation,

accompanied by the distortion in the models which we referred to in Section 6. A feature making for less delay even in the appearance of *He* II after a stellar brightening, at these highest excitations, compared with Z Andr and T Cor B, is the difference between radiation temperature and temperature of electron capture: this feature we have included in selecting the values in the models for Figs. 2, 3. In T Cor B the lag in all lines would be partly due to the lesser radiation temperature relative to capture temperature.

11. *Transition (RW Hydr) towards steady pair (α Scor).*—Since a major enquiry, towards which these problems of delay-time in fluorescence are one possible contribution, concerns the reason why the blue-red binary association should cause eruption, it is important to add to the more violent a case where the time-lag vanishes into steady state, and also a transition case between that limit and the eruptive binaries.

(*a*) RW Hydr, investigated by Merrill (**16**) and by Swings and Struve (**17**), has the following characteristics separating it from the other blue-red pairs.

(i) The amplitude, 9·7–10·9 photographic, is small compared with the others, although the period at about 370 days is not abnormal for M stars and Merrill groups the system among those showing lag of its nebular lines about $\pi/5$ behind phase.

(ii) Strength of *He* singlets greatly exceeds that of *He* triplets: this agrees with Z Andr but is opposite to the habit of AX Pers and planetary nebulae. In Swings and Struve's treatment this can accompany lower excitation and more line-stimulation of fluorescence than recombination.

(iii) On the other hand, although most of the *O* III spectrum resembles that of WC stars it is weaker than that of the other blue-red binaries in the 3760 A. of *O* III, which comes from line-excitation by *He* II. This tendency to recombination rather than line-excited fluorescence, indicating nebular rather than Wolf–Rayet conditions, is reconcilable with (ii), according to Swings and Struve, if the circumstellar gases are stratified; here as elsewhere those authors suggest that relaxation times may be considerable, and we have in our treatment given reason for associating stratification with the differences between times of relaxation. For we have shown (Section 6) that an optical appearance of strata can arise, during *changes* in the exciting star, from the differential rates of Grotrian's delay in fluorescent response to the irradiation.

In this star, however, the changes are less violent than in any of those previously discussed. In fact, besides the abnormally low amplitude of pulsation, a *slow* growth in excitation of the blue star is the slight survivor of the mechanisms which we listed in Section 2. For instance, the ratio *He* II/*He* I rose slowly between 1940 April and 1941 May, and the ratio *He* I/*Ca* II rose also by 1942. It seems more reasonable to attribute this to intrinsic progress in the source of excitation ((I) of Section 3) than to delays in response of Grotrian's type. There are, however, changes in the line velocities which may well be significant for the problem of Grotrian's *r*, which here becomes more complex than in any of our models. Superposed on the fact (shared with AX Pers, Z Andr and R Aqar) that forbidden nebular lines are in general more negative in radial velocity than *H*, the following are among changes recorded in some detail as to time. In 1937, from February 27 to April 30 and June 18, Balmer line velocities altered from +8 km./sec. to

+20; *He* I, on these dates, +10, +8, +28, with nebular lines +34, −13, −9. Balmer lines themselves changed from −5 to +42 in the longer interval 1932–5. At nebular densities it is not impossible for differential radiation pressure to separate the gases, so it is by no means certain to what extent these velocity changes record the appearance of luminosity at successive distances or successive stages of a re-adjustment of the kind investigated here. To any extent to which the lag in pulsation can be attributed to delay in atomic recombination, $\pi/5$ is about 70 days, a lag which in Fig. 1 for hydrogen would correspond to a distance of 10^{15} cm. from the radiating star, and for all other lines accessible to calculation the distance is less. If the observable changes, which are certainly faster than the slow *visible* development of the B star, involve Grotrian's mechanism, the two or three months which they have occasionally lasted will imply again about 10^{15} cm. for the radius of the fluorescing sphere.

Such an order of magnitude would eliminate some of the possibilities in the location of bright lines: interstellar nebulae at even planetary nebular distances are too large, and a Wolf–Rayet atmosphere far too small, and even a veil between the two stars is difficult to imagine large enough. The meagre evidence contributed by this interpretation of observed times favours therefore a large example of the location (iii), or a not too distant shell of location (v): this latter is the less convincing, since RW Hydr has had a less eruptive history so far as is known, and recent ex-nova shells cannot easily be invoked.

(*b*) α Scor is the limiting instance in lack of change; but it has interest not only for comparison but as the unique example near enough for dimensions to be directly accessible. Swings and Struve (**18**) establish to a high probability a diameter of [*Fe* II] nebulosity $5''{\cdot}9$ and separation of the B star by $3''$ ($3''{\cdot}2$, Adams and Joy (**19**)), from the M component, whose diameter has been estimated as 390○ (5×10^{13} cm.). Here, independently of any estimate of r from time of adjustment, and indeed in absence of any such evidence from so steady a pair, the location seems most probably between (i) and (v) for the bright lines; the latter is favoured by the fact that so large an extension (5×10^{15} cm.) would be rare for contemporary B ejecta, and is more easily attributable to a past which was not so quiet as the present. The problem of whether a B star can draw upon the atmosphere of its M companion is here likely to show its limiting possibilities: the extension of the M star, so far as limited to self-luminous area, places it just outside the B nebulosity, and the wide spacing of this particular visual binary reduces greatly the gravitational mingling of Kuiper's "contact" models, while the lack of large-scale pulsation also precludes any periodic interference between atmospheres which we listed as (iii), (iv). The excitation is also weaker than in the other stars, the nebular lines here being [*Fe* II] mainly, strengthened beyond the intensities of *Fe* II according to Struve (**20**) by a high density. It must be recollected that in M stars showing both *Fe* II and [*Fe* II] the latter sometimes strengthens relatively at minimum luminosity and smallest spread.

We suggest that these dimensions accord with a location of the nebular lines in old ejecta surrounding a now quiet star; if so, the under-luminous character of the B component, likened by Struve to the B component of Z Andr, may have a fresh significance. Instead of an intrinsic stellar property, it may be the effect of dimming by shell remnants. Z Andr, which has twice erupted in

the years when α Scor has remained quiet, has been screened by a shell which may also be responsible for its apparent under-luminosity, and as in planetary nebulae it is never certain that we see any of these stars undimmed. But in Z Andr there is no reason to suppose the M component is so far from the B as to lie outside the shell; the proximity in a smaller orbit may preclude so long a quiescence as that of α Scor, if the one component is the source of disturbance in the other.

The radial velocities in α Scor, +12·5 [*Fe* II] relative to −18·4 for *H* and *He* I absorption, are queried by Swings and Struve as not agreeing with the −0·1 km./sec. recorded for the emission lines by Wilson and Sanford. This is opposite, in either case, to the rule of the other blue-red binaries that velocities of forbidden lines are the more negative. The important [*Fe* II] lines, however, are stimulated by low voltage; of Struve's alternatives, they are probably due to electron impact in a density higher than that of P Cygn shells or most nebulae. This density would accord with the fact that the gas is no longer expanding now, compared with motion in the same lines in η Carinae and N Pict (1925); the density would also strengthen the opacity which we suggested as source of the under-luminosity in the embedded star.

Struve's hypothesis, that proximity of a red companion is essential to the eruptiveness of the blue star, thus passes to the limit where cause and effect both vanish, if low amplitude of pulsation gives the near quiescence of RW Hydr, and wide orbital separation gives the complete contemporary steadiness of α Scor: in neither case is there a pulsation sufficient for the major fluctuation in nebular targets which had exhibited the delays in fluorescence at their most striking.

12. *Conclusion.*—Existing data about the rates of spectroscopic change in six blue-red binaries are examined to find whether any appear to exemplify the times of delay in fluorescence calculated for various models by Grotrian's methods. The distinction between a near and quickly responding circumstellar gas, and a distant nebulosity surviving from earlier eruption and now slow in growth of its fluorescence, is reinforced from four of the binaries, by comparing delays in brightening with calculated rates of development in the spectra of models; location of the bright lines at various distances from the hotter star is thus made more understandable. Evidence for an intermediate location in the envelope of the M star is examined on the same principle, using time of development of the spectra. Comparison with the other two blue-red binaries correlates relative steadiness with greater distance of the M component, and slightness in eruptiveness of the blue star with low amplitude of pulsation of the M companion. The distinction we have drawn between inner and outer circumstellar nebulosity is a distinction in visibility, without necessarily implying maxima in density or composition; hence the conception of the inner shell as due to the binary association, at times concealing the outer nebulosity from ultra-violet irradiation, has promising significance for the problem of whether Wolf–Rayet binaries can similarly shield a planetary nebula from being stimulated to visible fluorescence, a problem we shall not pursue here.

Department of Physics,
University of Birmingham:
1951 *January* 3.

(1) I. S. Bowen, *Ap. J.*, **81,** 1, 1935.
(2) W. Grotrian, *Z. f. Astrophysik*, **13,** 215, 1937.
(3) P. Swings and O. Struve, *Ap. J.*, **93,** 356, 1941.
(4) G. Cillié, *M.N.*, **92,** 820, 1932.
(5) H. Zanstra, *Ap. J.*, **65,** 50, 1927.
(6) M. Johnson, *M.N.*, **110,** 84, 1950.
(7) D. B. McLaughlin, *P.A.S.P.*, **58,** 159, 1946.
(8) R. F. Sanford, *Ap. J.*, **109,** 81, 1949.
(9) C. Payne-Gaposchkin, *Ap. J.*, **104,** 357, 1946.
(10) P. W. Merrill, *Ap. J.*, **81,** 312, 1935.
(11) P. Swings and O. Struve, *Ap. J.*, **91,** 616–8, 1940.
(12) P. Swings and O. Struve, *Ap. J.*, **91,** 607–16, 1940.
(13) P. Swings and O. Struve, *Ap. J.*, **95,** 153, 1942.
(14) P. W. Merrill, *Ap. J.*, **99,** 15–17, 1944.
(15) C. Payne-Gaposchkin, *Ap. J.*, **104,** 368–9, 1946.
(16) P. W. Merrill, *Ap. J.*, **99,** 17–19, 1944 ; **111,** 484, 1950.
(17) P. Swings and O. Struve, *Ap. J.*, **94,** 296, 1941 ; **96,** 256, 1942.
(18) P. Swings and O. Struve, *Ap. J.*, **92,** 316, 1940.
(19) W. S. Adams and A. H. Joy, *P.A.S.P.*, **33,** 206, 1921.
(20) O. Struve, *Rev. Mod. Phys.*, **16,** 297, 1944.

A MODEL OF A SPIRAL GALAXY

G. C. McVittie and Cecilia Payne-Gaposchkin

(Received 1951 February 1)

Summary

It is assumed that the stars which constitute the spiral arms of a galaxy are formed from the disintegration of a bar of material which, at the moment of disintegration, is rotating like a rigid body. After their formation, the stars move under the Newtonian inverse-square-law attraction of a point-mass nucleus located at the centre of the bar. The stars move in elliptical orbits having their perigalactica or apogalactica on the initial position of the bar. The spiral arms are not orbits, but are curves defined by the " present positions " of the stars. The velocities to be expected are also found. The method of solution of the relevant equations is a graphical one. The principal conclusions are : (i) the shape of a galaxy alone does not determine its mass and the age of its arms, but shape, together with velocity of the stars in the arms, does so; (ii) arms may " lead " for parts of their extent and " trail " for the remainder; (iii) a fit can be obtained for the shapes of M51, M81 and M83 and for both shape and velocity of the outer parts of M33; the mass of M33 and the age of its arms are also found; (iv) the shape, mass and age of the Galaxy can be predicted and an explanation obtained of the " leading " arm in the Cygnus direction suggested by radio observations.

---

1. *Introduction.*—The aim of the investigation contained in this paper is to describe a simplified model of a spiral galaxy constructed on the basis of Newtonian mechanics. The initial configuration of the galaxy is envisaged as a cylindrical bar of material, of cross-section small compared with its length, at the centre of which is a spherical body, or nucleus, whose dimensions are also small compared with the length of the bar. The system is rotating slowly like a rigid body about an axis of fixed direction in space that passes through the centre of the nucleus. The main part of the mass of the system is supposed to be concentrated in the nucleus, the mass of the bar being negligibly small compared with that of the nucleus. At the initial "instant"* $t=0$, we imagine that the material of the bar produces stars which thereafter move in the gravitational field of the nucleus, the perturbations due to the mutual attractions between the stars themselves and the effects of the non-stellar remains of the bar being neglected. We have deliberately left aside the question of the origin of the bar which appears to us to be a separate problem. One may speculate that, before the formation of stars takes place, the bar consists of a mixture of dust and gas and the theory of its development out of some still more primitive configuration is one in which gravitational theory and the theories of hydrodynamics and of turbulence would presumably have to be employed. But it seems to us to be a plausible hypothesis that, when once the extra-nuclear stars have been formed, the motion of these objects will primarily be governed by the inverse-square-law attraction of the nucleus. The form of the spiral arms, in which the extra-nuclear stars are eventually distributed, should therefore be deducible from the elementary "planetary" theory of Newton if the initial

* For a definition of the initial " instant " see Section 5.

configuration is given. Very near to the nucleus, where the elementary potential function we employ cannot be expected to be valid, our results do not hold good.

The reader will find that the formation of the stars out of the material of the bar plays no part in our mathematical theory. All that we require is that the stars shall move in coplanar orbits, that the major axes of the orbits shall lie on a given line in space passing through the point-mass nucleus and that, initially, all the stars shall also be situated on this line.

We must emphasize that, in our theory, a spiral arm does not *consist* of the orbits of a set of stars. To borrow a term from E. A. Milne, the arm is *the locus of the "present positions"* of a set of stars whose orbits are, in our case, conic sections with a common major axis and with the nucleus at the focus. In this respect the theory traces its origins to a remark of S. Alexander*, and to the work of E. J. Wilczynski† and of E. A. Milne.‡ S. Alexander suggested that the break-up of a long thin cylinder of material might give rise to a spiral arm. Wilczynski showed that a set of stars moving in circular orbits round the nucleus, and starting from points along a common radius, would at a later time lie on a locus of spiral form. Finally E. A. Milne developed a theory analogous to ours except that he employed the gravitational theory of kinematical relativity instead of the classical theory of Newton. The investigations of E. W. Brown§ and of B. Lindblad|| fall into a different category, for they specify the motions of the stars through potential functions which are of a more elaborate type than the simple potential of a spherical mass which we employ.

2. *Position-curves.*—We consider first the problem of determining the position at time t of a star which left the bar at time $t=0$. We assume that the star moves in the gravitational field of a point mass M located at the centre of the bar, the gravitational field of the bar and of the star itself being neglected. We also assume that the velocity of the star is perpendicular to the bar at time $t=0$, which means that either the perigalacticon or the apogalacticon of the star's orbit lies on the line in space occupied by the bar at the initial instant.

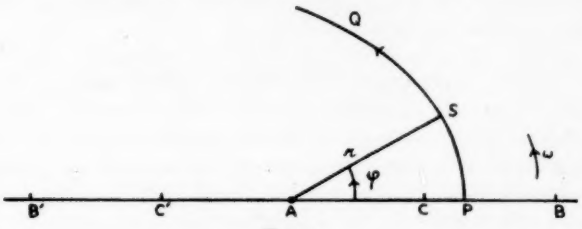


FIG. 1.

Fig. 1 shows the position of the bar B'B at time $t=0$, the central mass being at A. PQ is the orbit of a star which has perigalacticon at P and r, ϕ are the polar coordinates of the star S at time $t(>0)$. The orbit of the star is

$$\frac{l}{r} = 1 + e\cos\phi, \qquad (2.1)$$

* S. Alexander, *A.J.*, **2**, 95, 97, 105, 113, 126, 140, 148, 158, 1852. See in particular the article on p. 105.

† E. J. Wilczynski, *Ap. J.*, **4**, 97, 1896; *A.J.*, **20**, 67, 1899–1900.

‡ E. A. Milne, *M.N.*, **106**, 180, 1946.

§ E. W. Brown, *Ap. J.*, **61**, 111, 1925.

|| B. Lindblad, *Stockh. Obs. Ann.*, passim, 1936–1950.

and the relationship between ϕ and t is given by

$$\frac{d\phi}{dt} = \frac{h}{r^2}. \tag{2.2}$$

The semi-latus rectum, l, of the orbit is, of course,

$$l = \frac{h^2}{\gamma M}, \tag{2.3}$$

whilst the eccentricity, e, of the orbit may have any value from 0 to 1. Substituting in (2.2) from (2.1) and (2.3) we obtain, for the time t taken by the star to travel from P to S,

$$\left.\begin{aligned} t &= \frac{h^3}{\gamma^2 M^2}\int_0^{\phi}\frac{d\phi}{(1+e\cos\phi)^2} = \frac{h^3}{\gamma^2 M^2}\frac{F(e,\phi)}{(1-e^2)^{3/2}} \quad (0\leqslant e<1),\\ t &= \frac{h^3}{\gamma^2 M^2}\int_0^{\phi}\frac{d\phi}{(1+\cos\phi)^2} = \frac{h^3}{\gamma^2 M^2}F(1,\phi) \qquad (e=1), \end{aligned}\right\} \tag{2.4}$$

where the functions $F(e,\phi)$, $F(1,\phi)$ have the following forms*:—

$$\left.\begin{aligned} &\text{If } e=0, \qquad F(0,\phi)=\phi;\\ &\text{If } 0<e<1, \quad F(e,\phi)=2\tan^{-1}\left(\sqrt{\frac{1-e}{1+e}}\tan\frac{\phi}{2}\right)-e\sqrt{1-e^2}\,\frac{\sin\phi}{1+e\cos\phi};\\ &\text{If } e=1, \qquad F(1,\phi)=\tfrac{1}{2}\left(\tan\frac{\phi}{2}+\tfrac{1}{3}\tan^3\frac{\phi}{2}\right). \end{aligned}\right\} \tag{2.4a}$$

If we eliminate h from (2.4) and (2.1), (2.3), we find

$$\left.\begin{aligned} \gamma M t^2(1-e^2)^3 &= r^3(1+e\cos\phi)^3F^2(e,\phi) \quad (0\leqslant e<1),\\ \gamma M t^2 &= r^3(1+\cos\phi)^3F^2(1,\phi) \qquad (e=1), \end{aligned}\right\} \tag{2.5}$$

and we shall call the curves represented by these equations for given e and given Mt^2, *position-curves*. Obviously, the (r,ϕ) coordinates of S at time t are obtained for given l, e and Mt^2, by a simultaneous solution of equations (2.1) and (2.5). We shall adopt a graphical method of performing this operation, which will be described in more detail below.

At this stage in the investigation we have two alternatives open to us. One alternative—which we shall adopt—would be to assume that we had a set of stars which left different points of the bar at time $t=0$ and moved in the same sense in coplanar orbits of *different* eccentricities but with the bar as common major axis. At the end of time t we can determine, by the method described above, the coordinates of each star. The locus obtained by joining the positions of the stars we shall call a "*spiral arm*". A second alternative, however, is to assume that the stars leave the bar at $t=0$ and move as described above in orbits with *the same* eccentricity. The spiral arm and the position-curve, at a later time t, are identical in this case and we obtain indeed the generalization of Wilczynski's theory.† In particular, if the eccentricity is zero, as he assumes, the equations (2.5) and (2.4a) give

$$r^3\phi^2 = \gamma M t^2,$$

which is equivalent to his equation for a spiral arm.

* See, e. g., E. J. Routh, *A Treatise on Dynamics of a Particle*, C.U.P. 1898; Section 342 and Section 349.

† E. J. Wilczynski, *loc. cit.*

Adopting the first alternative, we must next consider the case when P is the apogalacticon of the star. The orbit is now

$$\frac{l}{r} = 1 + e\cos(\phi+\pi) = 1 - e\cos\phi, \tag{2.6}$$

whilst instead of (2.4) we have, writing $\psi = \phi + \pi$,

$$t = \frac{h^3}{\gamma^2 M^2}\int_{\pi}^{\phi+\pi} \frac{d\psi}{(1+e\cos\psi)^2} = \frac{h^3}{\gamma^2 M^2}\frac{F(e,\phi+\pi)-\pi}{(1-e^2)^{3/2}} \quad (0 \leqslant e < 1), \tag{2.7}$$

because, by (2.4 *a*), $F(e,\pi) = \pi$ for all e in $0 \leqslant e < 1$. The position-curve deduced from (2.6), (2.7) and (2.3) is

$$\gamma M t^2 (1-e^2)^3 = r^3(1-e\cos\phi)^3\{F(e,\phi+\pi)-\pi\}^2. \tag{2.8}$$

3. *The velocity.*—Consider again a star moving in an orbit of eccentricity e, and semi-latus rectum l, perigalacticon being at P. The magnitude of the velocity v of the star at S is, of course, given by

$$v^2 = \gamma M\left(\frac{2}{r} - \frac{1-e^2}{l}\right). \tag{3.1}$$

Using (2.1), (2.3), we have

$$v^2 = \frac{\gamma^2 M^2}{h^2}\{1 + 2e\cos\phi + e^2\}. \tag{3.2}$$

If we eliminate h from (3.2), (2.4), we find

$$\left.\begin{aligned} v^2 &= \left(\frac{\gamma M}{t}\right)^{2/3}\frac{1+2e\cos\phi+e^2}{1-e^2}F^{2/3}(e,\phi) \quad (0 \leqslant e < 1), \\ v^2 &= 2\left(\frac{\gamma M}{t}\right)^{2/3}(1+\cos\phi)F^{2/3}(1,\phi) \qquad (e = 1). \end{aligned}\right\} \tag{3.3}$$

Hence, for each point on a spiral arm, the corresponding velocity can be found by inserting the values of e, ϕ for the point on the spiral arm into formulae (3.3).

If P is the apogalacticon of the star's orbit and $0 \leqslant e < 1$, the formula for the velocity reads, by (2.6), (2.7) and (3.2),

$$v^2 = \left(\frac{\gamma M}{t}\right)^{2/3}\frac{1-2e\cos\phi+e^2}{1-e^2}\{F(e,\phi+\pi)-\pi\}^{2/3}. \tag{3.4}$$

4. *The eccentricities and semi-latera recta.*—The foregoing formulae apply to spiral arms formed by the motion of a set of stars whose orbits are restricted only by the assumptions that (*a*) the orbits are coplanar and are traced out in the same sense, (*b*) the orbits have either their perigalactica or apogalactica on the bar B′B at time $t = 0$. We shall now restrict the motion further by assuming that, at $t = 0$, the bar was rotating like a rigid body with angular velocity ω and that each star had, initially, the velocity of the point on the bar at which the star left it. In Fig. 1 let $AP = r_i$ and let the star's velocity then be v_i. It therefore follows that, in the star's orbit,

$$h = r_i v_i. \tag{4.1}$$

Suppose further that P is the perigalacticon of the star's orbit. By (3.2) and (4.1) we have

$$v_i^2 = \frac{\gamma^2 M^2}{r_i^2 v_i^2}(1+e)^2,$$

so that

$$v_i^2 = \frac{\gamma M}{r_i}(1+e). \tag{4.2}$$

But now

$$v_i = r_i\omega, \tag{4.3}$$

so that (4.2) yields the two equations

$$1+e = \frac{r_i^3\omega^2}{\gamma M} = \frac{v_i^3}{\gamma M\omega}. \tag{4.4}$$

On the other hand, if P is the apogalacticon of the star's orbit, we have, instead of (4.2) and (4.4),

$$v_i^2 = \frac{\gamma M}{r_i}(1-e) \tag{4.5}$$

and

$$1-e = \frac{r_i^3\omega^2}{\gamma M} = \frac{v_i^3}{\gamma M\omega}. \tag{4.6}$$

Obviously there will be a point C on the bar for which the star's orbit will be a circle with centre at A. The radius of this orbit is r_c and the corresponding circular velocity is v_c where, by (4.4) or (4.6),

$$1 = \frac{r_c^3\omega^2}{\gamma M} = \frac{v_c^3}{\gamma M\omega}. \tag{4.7}$$

Again a star with a parabolic orbit starts from a point B on the bar, where $AB = r_p$, and it has a velocity v_p. These quantities are found from (4.4) with $e=1$ and satisfy

$$2 = \frac{r_p^3\omega^2}{\gamma M} = \frac{v_p^3}{\gamma M\omega}. \tag{4.8}$$

Examination of the formulae (4.4) to (4.8) shows that the stars whose perigalactica lie on the bar start from points lying between C and B, whilst those whose apogalactica lie on the bar start from points between A and C.

Since $l = h^2/(\gamma M) = r_i^2 v_i^2/(\gamma M)$, the formulae (4.4) and (4.6), when combined with (4.7), yield

$$\left.\begin{aligned} l &= r_c(1+e)^{4/3} \quad \text{(perigalacticon on the bar)},\\ l &= r_c(1-e)^{4/3} \quad \text{(apogalacticon on the bar)}, \end{aligned}\right\} \tag{4.9}$$

Hence the equations of the orbits, starting from perigalacticon and apogalacticon, respectively, are now, by (2.1) and (2.6),

$$\frac{r_c}{r} = \frac{1+e\cos\phi}{(1+e)^{4/3}} \tag{4.10}$$

and

$$\frac{r_c}{r} = \frac{1-e\cos\phi}{(1-e)^{4/3}}. \tag{4.11}$$

Consideration of the orbits with perigalactica on AC indicates that the formation of stars from the bar cannot be supposed to occur at points too close to $r=0$. An orbit of eccentricity e whose apogalacticon lies on AC will have its perigalacticon on AC′ at a distance $(1-e)^{4/3}r_c/(1+e)$ from A. The velocity at the apogalacticon is, by (3.1), (4.9),

$$\left(\frac{\gamma M}{r_c}\right)^{1/2}(1-e)^{1/3},$$

which is, of course, less than

$$v_c = \left(\frac{\gamma M}{r_c}\right)^{1/2},$$

but at the perigalacticon of the orbit the velocity is

$$\left(\frac{\gamma M}{r_c}\right)^{1/2} \frac{1+e}{(1-e)^{2/3}},$$

which is not only greater than v_c but tends to infinity as e tends to 1. The reason for this is that we have idealized the nucleus into a point-mass at A, whereas in nature it will extend to a considerable distance from this point. Hence the bar does not begin until this distance is exceeded, and our theory therefore cannot be applied up to the neighbourhood of the point $r=0$.

5. *Graphical methods.*—We must now adapt the equations to a form suitable for obtaining a graphical solution of (4.10), (2.5) and of (4.11), (2.8). We shall use, as our fundamental units, the kiloparsec for r, the solar mass for M, and the year for t. This will mean that the velocity v should be expressed in Kps./year but, as this is an inconvenient unit for velocity, we shall retain the usual expression of velocity in km./sec. and write $v=\alpha V$, where α is the conversion factor and V is in km./sec. If we denote by $K, M_\odot, Y$, respectively, the kiloparsec, solar mass and year, we have

$$\left.\begin{aligned} &\gamma = 6{\cdot}670\times 10^{-8}\,\text{cm.}^3\,\text{g.}^{-1}\,\text{sec.}^{-2} = 4{\cdot}491\times 10^{-24} K^3 M_\odot^{-1} Y^{-2}, \\ &\text{whilst} \\ &\alpha = 1{\cdot}022\times 10^{-9}. \end{aligned}\right\} \qquad (5.1)$$

In order to plot a fundamental set of curves we choose

$$M = M_0 = 10^{10} M_\odot, \quad t = t_0 = 10^8 Y, \quad r_c = (r_c)_0 = 2{\cdot}5 \text{ Kps.} \qquad (5.2)$$

Using (5.1) we obtain for the constants that will appear in the equations of the fundamental curves the values

$$\left.\begin{aligned} &\tfrac{1}{3}\log_{10}\gamma M_0 t_0^2 = 0{\cdot}884, \qquad \log_{10} 2{\cdot}5 = 0{\cdot}398, \\ &\tfrac{1}{3}\log_{10}\frac{\gamma M_0}{t_0} - \log_{10}\alpha = 1{\cdot}874. \end{aligned}\right\} \qquad (5.3)$$

The fundamental position-curves for the stars which leave the bar at points on CB are, by (2.5), (5.3),

$$\log r = 0{\cdot}884 - \tfrac{1}{3}\log\frac{(1+e\cos\phi)^3 F^2(e,\phi)}{(1-e^2)^3} \quad (0\leqslant e<1), \qquad (5.4)$$

$$\log r = 0{\cdot}884 - \tfrac{1}{3}\log\{(1+\cos\phi)^3 F^2(1,\phi)\} \quad (e=1), \qquad (5.5)$$

whilst the fundamental orbits by (4.10), (5.3) are

$$\log r = 0{\cdot}398 - \log\frac{1+e\cos\phi}{(1+e)^{4/3}} \quad (0\leqslant e\leqslant 1). \qquad (5.6)$$

The corresponding fundamental velocity curves are, by (3.3), (5.3),

$$\log V = 1{\cdot}874 + \tfrac{1}{2}\log\left\{\frac{1+2e\cos\phi+e^2}{1-e^2}\,F^{2/3}(e,\phi)\right\} \quad (0\leqslant e<1), \qquad (5.7)$$

$$\log V = 1{\cdot}874 + \tfrac{1}{2}\log\{2(1+\cos\phi)F^{2/3}(1,\phi)\} \quad (e=1). \qquad (5.8)$$

The fundamental curves for stars which have their apogalactica on AC are

$$\log r = 0{\cdot}884 - \tfrac{1}{3}\log\left[\frac{(1-e\cos\phi)^3\{F(e,\phi+\pi)-\pi\}^2}{(1-e^2)^3}\right], \tag{5.4a}$$

$$\log r = 0{\cdot}398 - \log\frac{1-e\cos\phi}{(1-e)^{4/3}}, \tag{5.6a}$$

$$\log V = 1{\cdot}874 + \tfrac{1}{2}\log\left[\frac{1-2e\cos\phi+e^2}{1-e^2}\{F(e,\phi+\pi)-\pi\}^{2/3}\right]. \tag{5.7a}$$

If, in Fig. 1, AC′ = AC and C′B′ = CB, the fundamental curves for stars which leave the bar between A and C′ and between C′ and B′ may be obtained from those given above by rotating the diagrams clockwise through π.

Suppose now that we wish to obtain the position-curves, orbits and velocity-curves for a different initial choice of M, t and r_c. We write

$$M = \mu M_0, \quad t = \tau t_0, \quad r_c = \sigma(r_c)_0. \tag{5.9}$$

Inspection of the equations (2.5) and (2.8), (4.10) and (4.11), and (3.3) and (3.4), shows that we can obtain the new curves from the fundamental ones by making the following alterations in equations (5.4) to (5.7 *a*), for the position-curves, the orbits and the velocity-curves, respectively:

$$\left.\begin{array}{lll} \text{In (5.4), (5.5), (5.4}a\text{)} & \text{replace} & 0{\cdot}884 \text{ by } 0{\cdot}884 + \tfrac{1}{3}\log\mu\tau^2; \\ \text{,, (5.6), (5.6}a\text{)} & \text{,,} & 0{\cdot}398 \text{ ,, } 0{\cdot}398 + \log\sigma; \\ \text{,, (5.7), (5.8), (5.7}a\text{)} & \text{,,} & 1{\cdot}874 \text{ ,, } 1{\cdot}874 + \tfrac{1}{3}\log\dfrac{\mu}{\tau}. \end{array}\right\} \tag{5.10}$$

The interpretation of the factors σ, μ in terms of the new angular velocity of the bar is of interest. Let ω_0 be the angular velocity of the bar for the choice $M = M_0$ and $(r_c) = (r_c)_0$, so that, by (4.7),

$$(r_c)_0 = \left(\frac{\gamma M_0}{\omega_0^2}\right)^{1/3}, \quad (v_c)_0^2 = (\gamma M_0\omega_0)^{1/3}.$$

Let ω be the angular velocity of the bar for the choice (5.9) of fundamental constants. We have

$$r_c = \sigma(r_c)_0 = \left(\frac{\gamma\mu M_0}{\omega^2}\right)^{1/3} = \left(\frac{\mu\omega_0^2}{\omega^2}\right)^{1/3}(r_c)_0.$$

Hence

$$\omega = \frac{\mu^{1/2}}{\sigma^{3/2}}\omega_0, \tag{5.11}$$

whilst the new circular velocity is

$$v_c = \left(\frac{\mu}{\sigma}\right)^{1/2}(v_c)_0. \tag{5.12}$$

Finally we note that the fundamental angular velocity of the bar, ω_0, and the fundamental circular velocity, $(V_c)_0$, are, by (4.7), (5.1), (5.2) respectively,

$$\left.\begin{array}{l} \omega_0 = \left\{\dfrac{\gamma M_0}{(2{\cdot}5)^3}\right\}^{1/2} = 5{\cdot}362 \times 10^{-8} \text{ rad./}Y, \\ (V_c)_0 = \dfrac{(\gamma M_0\omega_0)^{1/3}}{\alpha} = 131 \text{ km./sec.} \end{array}\right\} \tag{5.13}$$

At this rate of rotation, the bar turns through about 3° in 10^6 years. Thus the "instant" $t = 0$ of the preceding calculations may in fact be an interval of

time as large as 10^8 years for, during that period, the position of the bar in space does not greatly alter.

6. *The graphs.*—The graphical analysis of the preceding theory will now be described. Fig. 2 shows the position-curves given by the equations (2.5) for the special case of circular orbits ($e=0$); this is the case discussed by Wilczynski.* The curves are drawn for a mass of $10^{10}M_\odot$ and for several values of t, as

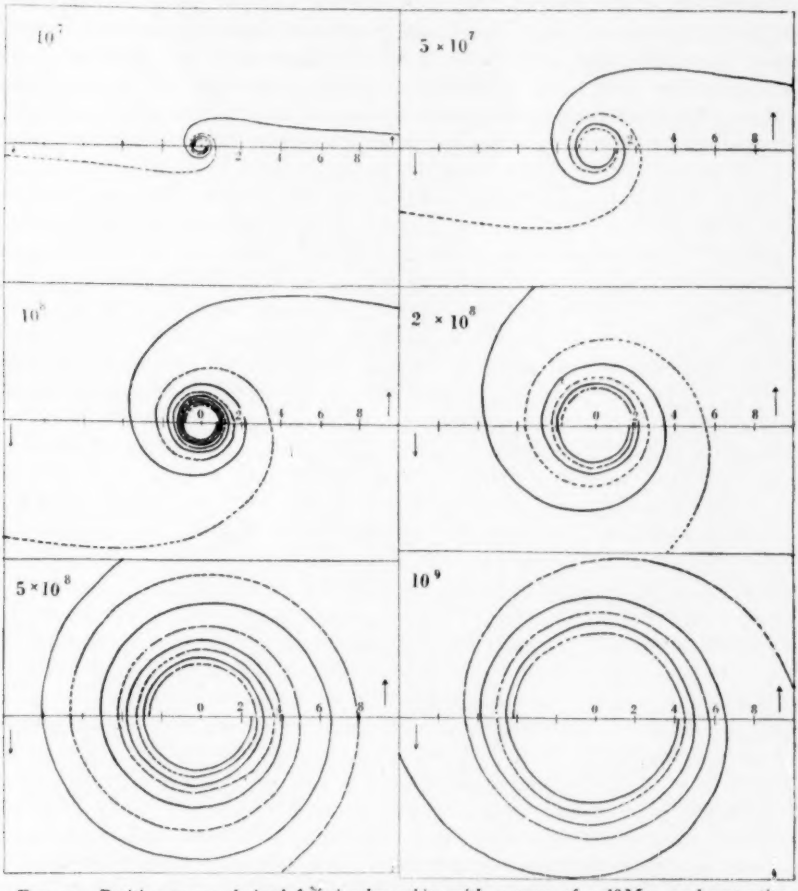


FIG. 2.—*Position-curves derived for circular orbits with a mass of $10^{10}M_\odot$ and respective ages 10^7, 5×10^7, 10^8, 2×10^8, 5×10^8 and 10^9 years. The arrows beside the line denoting the zero-point of ϕ indicate the direction of orbital motion and of increasing ϕ. The radial scale is marked in Kps. The first two curves are carried through 4π radians, the third and fifth through 8π, the fourth and sixth through 6π. All the curves, of course, undergo an infinite number of turns before they reach the central point.*

provided for by (5.10). Examination of the curves will illustrate: (*a*) the progressive winding-up of the position-curves (which in this case are identical with the spiral arms) with advancing age; (*b*) the gradual expansion of the tightly wound portions of the spiral with increasing t; and (*c*) the outermost portions of the arms, approaching asymptotically the line corresponding to $\phi=0$.

* E. J. Wilczynski, *loc. cit.*

Fig. 3 illustrates the form of the position-curves derived from (2.5), with various eccentricities, for our fundamental case (5.2). Evidently the curves become more and more distorted, as compared with the curves for circular orbits (Fig. 2), the more the eccentricity deviates from zero. The curve for the largest eccentricity shown ($e=1$) reaches a limit at $\phi=\pi$, and the curves for eccentricities a little less than 1 approach it progressively. Curves for eccentricities greater than 1 (hyperbolic orbits) would obviously lie outside the parabolic curve, but we have not considered them, as stars in hyperbolic orbits probably play little part in the configuration of galaxies. The position-curves for stars in elliptical orbits starting from perigalacticon approach the nucleus most closely at values of ϕ near to $2n\pi$; those for stars in elliptical orbits starting from apogalacticon, at values of ϕ near to $(2n-1)\pi$.

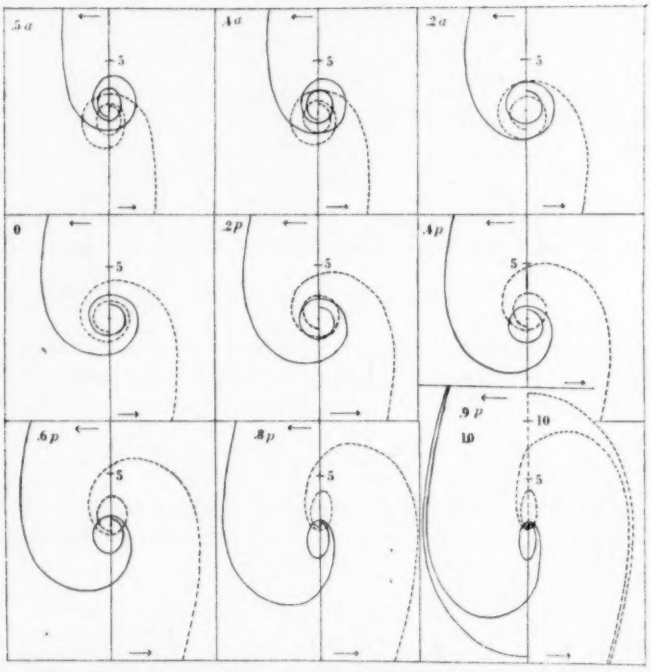


FIG. 3.—*Position-curves derived for orbits of various eccentricities (as indicated in the diagram) with a mass of* $10^{10}M_{\odot}$ *and a time of* 10^{8} *years since the liberation of the stars from the bar. a denotes apogalacticon start; p, perigalacticon start. Arrangement is as in Fig. 2. All the curves except the first two are carried through* 4π *radians; the first two are carried through* 5π *radians.*

Our method for finding the position at any time of the spiral arm occupied by the stars liberated from a rigidly rotating bar (Section 4) involves determining the points of intersection of the position-curves of Fig. 3 with the orbit-curves (4.10), (4.11). The graphical determination of the intersections of the position-curves and the orbit-curves that correspond to the same eccentricity and starting-orientation is most easily performed if the logarithms of the distances are plotted against ϕ. By this expedient it is possible to use one set of curves for reading off the intersections that correspond to any chosen

combination of mass, time and σ (see (5.10)), by the addition of suitable constants to the log r-values for the position-curves and orbit-curves.

Fig. 4 shows the spiral arms that are produced by the loci of these intersections for the case where $M=10^{10}M_{\odot}$ and for various values of the time. As in Fig. 2, the noteworthy features of the curves are: (*a*) the progressive winding-up of the spiral arm with advancing age; (*b*) the gradual expansion of the closely wound portion; (*c*) the existence of two extended arms (which are no longer asymptotic to the zero direction of ϕ, except for values of t smaller than those shown in the figure). Several features that are not shared by the special case of circular orbits are worthy of note. The arms now cross one another at intervals of about π in ϕ; the crossing-points are close to the nucleus for small values of t, but move outward progressively with advancing age. Furthermore, the arms are distorted by the effects of the shape of the position-curves for eccentric orbits, and the distortion is such that, in every circuit within the first intersection, the arm is "leading" (i.e. receding from the nucleus with increasing ϕ) through about a half-turn, and "trailing" (i.e. approaching the nucleus with increasing ϕ) for the other half. This effect becomes more and more conspicuous the nearer the nucleus is approached. Finally, if the parabolic orbits are regarded as marking the effective outer limit of the model galaxy, the size, and therefore the apparent length of the arms, increases progressively with age.

The orbital velocities of the stars at any time (equations (3.3), (3.4)) increase progressively, in the case of circular orbits, with increasing ϕ and hence with increasing r. For any eccentricity other than zero, this is no longer true. To obtain a complete picture of the velocity-curves it is necessary to examine the velocity as a function of ϕ rather than of r (which latter is the usual procedure in dealing with observations of the rotational velocities of galaxies), since both the individual position-curves, and also the deduced spiral arms, pass through the same value of r more than once with different values of ϕ. When the velocities deduced from our equations are plotted against r, they turn back on themselves repeatedly inside the point at which the arms begin to intersect.

Fig. 5 shows the course of several velocity-curves when V is plotted as ordinate against ϕ as abscissa. These curves are obtained by reading V from equations (5.7), (5.7*a*) and (5.8) at the values of ϕ given by the spiral arms for the appropriate eccentricities. All the curves are computed for $\mu/\tau=1$. On the other hand, when V is regarded as a function of r, and a V-r curve plotted, it is evident that the mean velocities taken over a range of r will often cover more than one section of the curve, except for the outermost regions of the spiral, because several values of V may correspond to one value of r. When V is plotted against ϕ, however, each curve attains any given value of the abscissa only once.

7. *Comparison with observation.*—When the sequence of model spirals (Fig. 4) is compared with the actual galaxies that show spiral form, they will be found to reproduce many of the observed features fairly well. Fig. 4 shows the form and dimensions of a spiral for $M=10^{10}M_{\odot}$ and $r_c=2{\cdot}5$ Kps. at several intervals since the initial instant. A "young" spiral thus consists of two outer trailing arms, which begin to cross and to "lead" periodically, at a relatively small distance from the nucleus. As its age advances, the length of the outer arms and the distance from the nucleus at which the arms begin to cross, increase progressively. The effect is that structure is obliterated in a central

region that grows in size with increasing age. If this progression is compared with the series of normal spirals, our model galaxy passes with advancing age through forms that resemble those of Sc, Sb and Sa spirals.*

We have already mentioned in Section 4 that our theory cannot be expected to reproduce the true state of affairs after we pass to within a certain distance of the nucleus. If this distance is larger in some systems than in others (which will depend among other things on the distribution of mass) the regions near

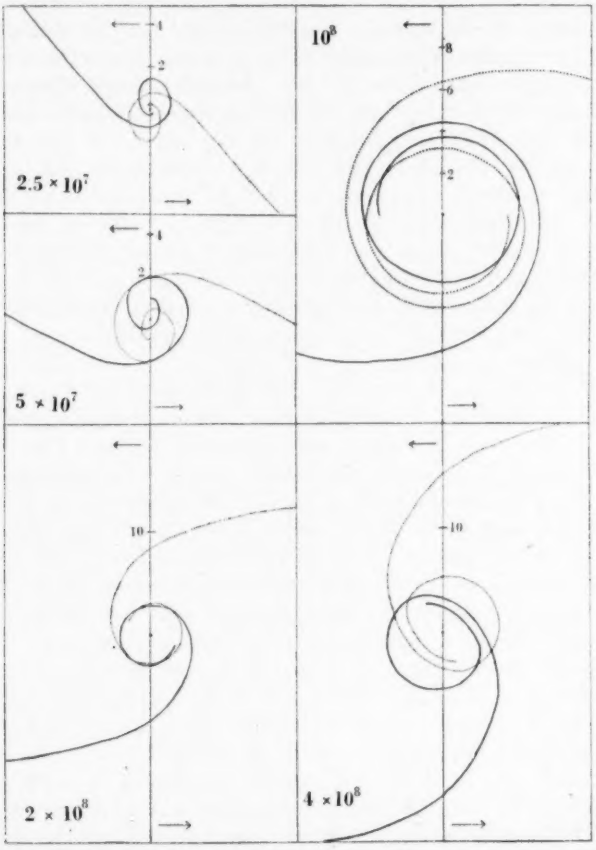


FIG. 4.—*Forms of spiral arms derived from intersections of position-curves* (*Fig.* 3) *and orbit-curves for various eccentricities. The curves are obtained for* $M = 10^{10} M_{\odot}$, $r = 2{\cdot}5$ *Kps. and for various ages. The radial scale is marked in Kps.; note that the scale is halved for the two last curves.*

the nucleus may tend to move more nearly with rigid rotation. That this is true at some distance in all galaxies is shown by the measured radial velocities. Possibly the barred spirals fall into this category where the "rigid" region is larger than for normal spirals and may even contain the inner portions of the bar in an "undisintegrated" state.

For our values of the mass and of r_c (corresponding to $\mu = 1$, $\tau = 1$, $\sigma = 1$ in (5.10)), the age of the model galaxy is determinate. But we can reproduce the

* E. P. Hubble, *The Realm of the Nebulae*, Yale Univ. Press, 1936, p. 45.

same sequence of forms and dimensions by choosing values for some or all of μ, τ, σ other than unity. As a first example, if $\sigma = 1$ but μ, τ have any pair of values satisfying $\mu\tau^2 = 1$, the same sequence results but $\tau = \mu^{-1/2}$. In this case, also, by (5.11), (5.12) and the definition of r_c, we have

$$\omega = \mu^{1/2}\omega_0, \qquad v_c = \mu^{1/2}(v_c)_0, \qquad r_c = (r_c)_0.$$

As a second example, suppose that μ, τ and σ satisfy the two equations $\mu\tau^2 = \sigma^3$ and $\mu = \sigma^3$. By (5.10), the effect of the first relation is to move the origins of $\log r$ in the position- and orbit-curves through the same amount, so that the same sequence again results. But now $\tau = 1$ and

$$\omega = \omega_0, \qquad v_c = \mu^{1/3}(v_c)_0, \qquad r_c = \mu^{1/3}(r_c)_0.$$

As a third example, which again gives the same sequence, let μ, τ and σ satisfy the two equations $\mu\tau^2 = \sigma^3$ and $\mu = \sigma$. Hence $\tau = \mu$ and

$$\omega = \mu^{-1}\omega_0, \qquad v_c = (v_c)_0, \qquad r_c = \mu(r_c)_0.$$

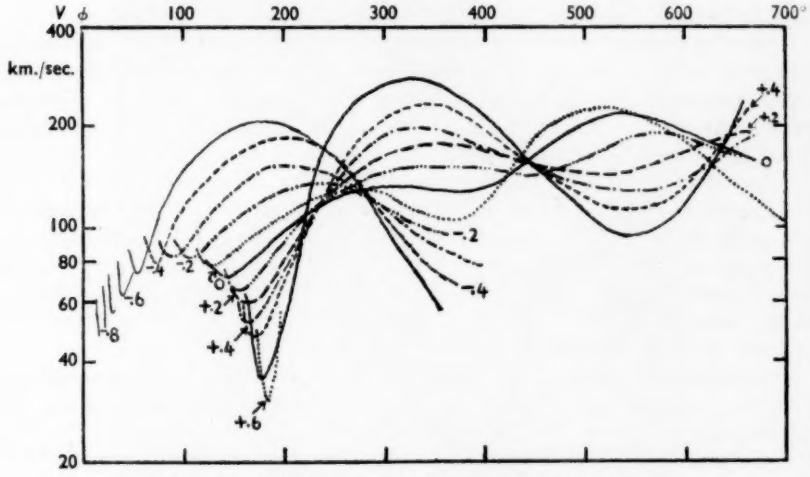


FIG. 5.—*Velocity-curves for various values of* $\frac{1}{3}\log \mu\tau^2$.

Thus in all three cases the galaxy passes through the same sequence of forms and dimensions but at rates which, in the first and third cases, depend on the mass. It also follows that position in the sequence does not determine the age of a galaxy uniquely: it merely indicates some appropriate combination of the values of μ, τ, σ.

But while the same sequence of forms and sizes can thus be produced by suitable combinations of μ, τ and σ, the corresponding velocities will not be the same, since the velocity varies as $(\mu/\tau)^{1/3}$. Thus, although from a consideration of form and dimensions alone the age of the galaxy cannot be found, the rotational velocities make it possible to draw some conclusions about age and mass separately.

Comparison with individual galaxies is complicated by the great difficulty of allowing for the inclination of the system in the line of sight, and also in many cases by the uncertainty in the precise dimensions. Messier 51 appears to represent a system that is seen in plan, and it compares well with the third

diagram of Fig. 4, if i is taken equal to 27°. The inner arms are reproduced if constant eccentricity of 0·15 for the orbits, and apogalacticon starts, are assumed. Messier 83 also appears to be of roughly circular outline, and it corresponds to a greater age, as shown in the fifth diagram of Fig. 4. In this case we have taken $i=45°$. We have also succeeded in reproducing the arms of Messier 81, taking $i=60°$.

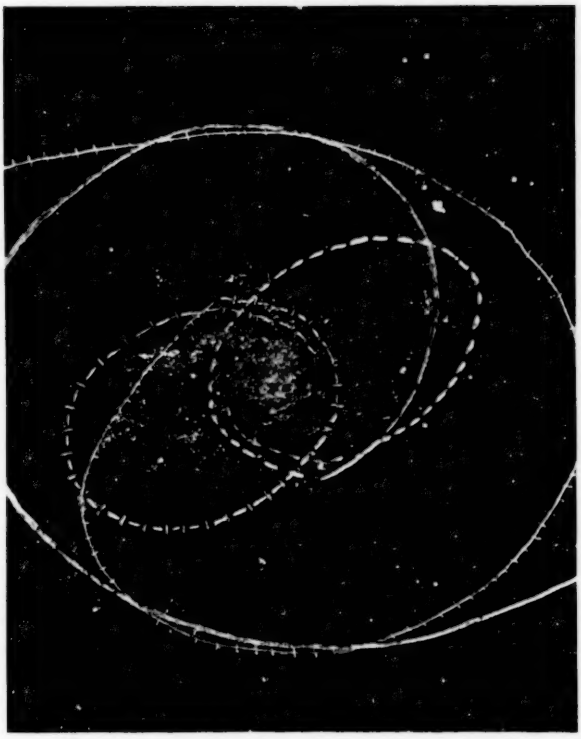
8. *Messier* 33.—The galaxy for which the available data are the most complete is M33, where the radial velocities have been studied by N. U. Mayall and L. H. Aller.* The data on velocities have been interpreted by A. B. Wyse and N. U. Mayall† in terms of a thin-disk model of a galaxy; our object is to show that these same data can be interpreted in terms of our model where the main mass is concentrated at the centre. Taking 220,000 parsecs‡ as the distance of M33, it follows that $1°=3·85$ Kps., and that the readily observable radius, within which lie all the points whose velocities were measured, is about two kiloparsecs. Mayall and Aller considered that the system is inclined at 33° to the line of sight, and we adopt this value, as well as the azimuths and true nuclear distances that are calculated from it.

The true nuclear distances (converted to Kps.) were plotted against the azimuths, and the resulting curve was compared with the curves of Fig. 6 which represents spiral arms drawn with $\log r$ and ϕ as ordinate and abscissa. Greatest weight was given to the outermost observable parts of the arms. It was judged that the curve corresponding to $M=10^{10}M_\odot$ (i. e. $\mu=1$), $r_c=2·5$ Kps. (i. e. $\sigma=1$) and $\frac{1}{3}\log\mu\tau^2=-0·20$, agreed best in form with the observed curve. For this curve it turns out that $r=3·75$ Kps. at $\phi=100°$. But the true distance of a point on the arm from the nucleus at $\phi=100°$ is 0·78 Kps. so that $\sigma=0·21$ or $\log\sigma=-0·682$. Hence the value of $\frac{1}{3}\log\mu\tau^2$ that will reproduce the outer arms in both shape and size is $-0·20-0·682=-0·882$. This keeps the difference of the zero points of the position-curves and orbit-curves for $\frac{1}{3}\log\mu\tau^2=-0·882$, $\sigma=0·21$, the same as the difference when $\frac{1}{3}\log\mu\tau^2=-0·20$, $\sigma=1$.

It remains to compare the observed rotational velocities with those predicted from our formulae. All the published velocities are of bright-line objects with nebular characteristics, and our theory applies specifically to stars. However, it seems probable that all these nebulae are associated with stars, and it is worth examining their velocities on the assumption that they share the stellar velocities, at least approximately.

We select from Fig. 5 the V-ϕ curve that corresponds to the spiral arm obtained for $\frac{1}{3}\log\mu\tau^2=-0·20$, and determine the change in $\log V$ that will produce the best fit with the observed values. Readings for ϕ were made from Plate 9 by associating each point for which a rotational velocity was available with the arm that runs nearest to it. The best fit with the V-ϕ curve is obtained when the logarithms of the velocities are decreased by 0·128, which corresponds to $\log(\mu/\tau)=-0·384$. But from the spiral arms corrected for scale we have obtained $\log(\mu\tau^2)=-2·646$, and so solving for μ and τ, we find that Messier 33 has a mass of $7·3\times10^8M_\odot$ and its arms have an age of $1·8\times10^7$ years. Mayall and Aller tabulate a mass of $1·7\times10^9M_\odot$, whilst Hubble estimates it at $1·5\times10^{10}M_\odot$, but this high value is ascribed by Mayall and Aller to the neglect

* N. U. Mayall and L. H. Aller, *Ap. J.*, **95**, 5, 1942.
† A. B. Wyse and N. U. Mayall, *Ap. J.*, **95**, 24, 1942.
‡ E. P. Hubble, *loc. cit.*, p. 143.



Predicted curves for Messier 33. The inner broken curves represent constant eccentricity 0·64, apogalacticon start.

of the velocity of the system. Mayall and Aller give ***rotational periods*** lying between 2×10^8 years (outer arms) and $5{\cdot}9 \times 10^7$ years (main body), both of these values, of course, predicating circular orbits.

It is seen from Fig. 7 that the predicted velocity-curve deviates greatly from the observed velocities for values of ϕ greater than about 400°. The predicted course of the arms also fails noticeably to reproduce the form of the system at about this point, which evidently marks the limit of applicability of our theory. This point corresponds to an orbit, starting from apogalacticon, with an

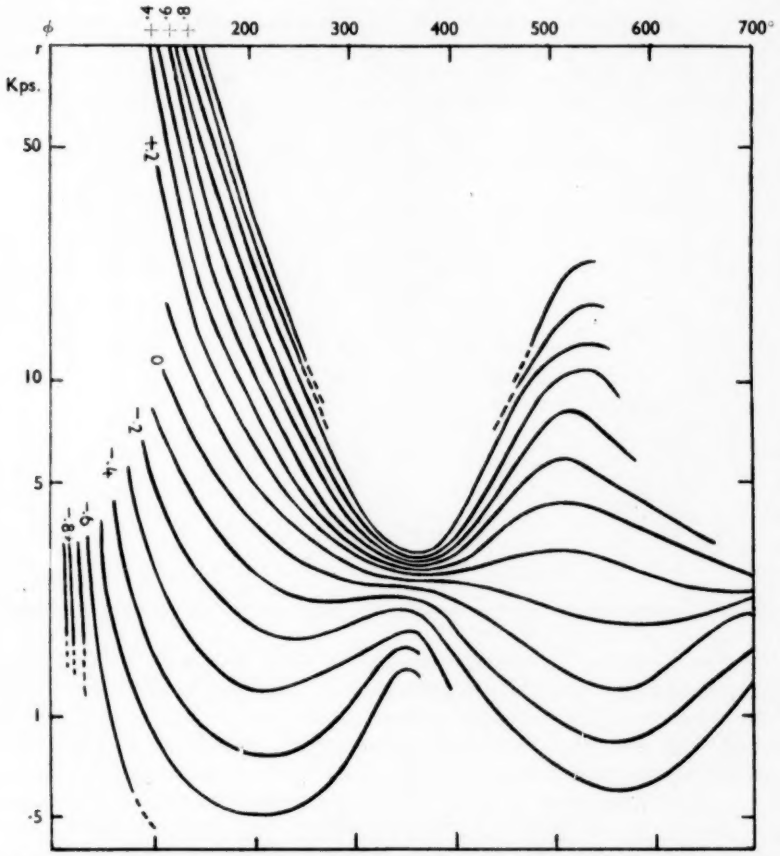


FIG. 6.—*Spiral arms for various values of* $\frac{1}{3}$ *log* μr^2 *plotted with log* r *and* ϕ *as ordinate and abscissa.*

eccentricity between 0·6 and 0·7. Suppose therefore that at this distance from the nucleus we abandon our hypothesis that the stars were produced from a rigidly rotating bar and replace it by the assumption that the bar breaks up into a set of stars which move in orbits of the *same eccentricity*. If we choose this eccentricity to be 0·64 and also assume that all the stars have their apogalactica on the bar, the spiral arms obtained—which are now identical with the position-curves—are indicated by the broken lines on Plate 9. It appears that the

spiral arms continued towards the nucleus in this way are in good agreement with observation. The spiral arm for constant e ($=0{\cdot}64$) has been carried through rather more than 2π in ϕ. It appears to enclose the bright inner portion of the observed arms, and another turn through 2π would evidently bring it near to the conspicuous inner arm. The velocities that would correspond to this calculated spiral arm at these values of ϕ would, however, fluctuate between the limits of 60 and 250 km./sec. whilst the velocities observed by Mayall and Aller lie, on the average, below the lower of these values, as may be seen from Fig. 7.

Thus it seems as if the outer arms of M33 can have been produced by the disintegration of a bar rotating like a rigid body but that the inner arms are produced by the disintegration of the bar into stars which all move in orbits of the same eccentricity, irrespective of their distances from the nucleus.

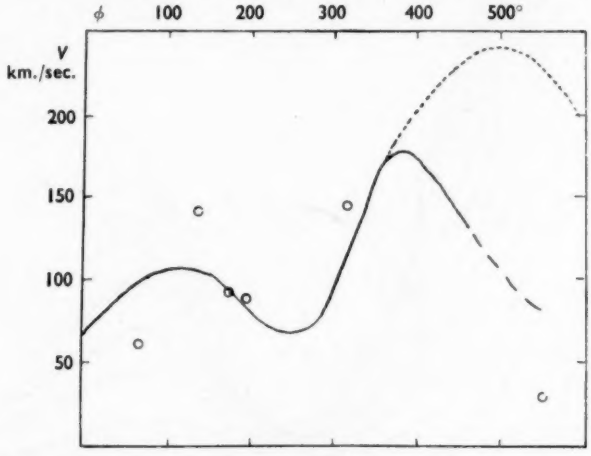


FIG. 7.—*Predicted velocity curve for Messier* 33. *Circles represent means of observed velocities.*

9. *The Galaxy.*—To select a model that shall represent the Galaxy is a different problem from a fitting of position-curves and velocity-curves to a system of known form. The tentative picture of the Galaxy shown in Fig. 8 was selected with the purpose of taking account of the following facts: (*a*) the existence of two condensations of stars (the "Oort condensations") which are about 2 Kps. apart, and may reasonably be supposed to represent sections of spiral arms; (*b*) the known direction (galactic longitude 327°) and distance (8 to 10 Kps.) of the galactic centre; (*c*) the brilliance of the Carina regions, where we are presumably looking along an arm; (*d*) the location of an intense radio maximum* at galactic longitude 57° differing by about 90° from the galactic centre in longitude; and (*e*) the solar velocity of about 250 km./sec., directed towards Cygnus ("galactic rotation"). The existence of the radio maximum at 90° from the galactic centre, in the direction towards which the Sun's rotational motion is observed to be, is interpreted by the radio observers as indicating that the arms in that direction may be "leading".

* J. G. Bolton and K. C. Westfold, *Nature*, **165**, 487, 1950.

The configuration shown in Fig. 8 has been obtained by trial and error and corresponds to one or other of the following values of the constants, according to the postulated distance of the Sun from the galactic centre:

Distance of Sun from galactic centre (Kps.)	σ	ω/ω_0	Time since initial instant (10^8 years)	Mass of Galaxy in $10^{11} M_\odot$
8	3·63	0·54	7·3	1·4
9	4·09	0·48	8·2	1·6
10	4·54	0·43	9·1	1·8

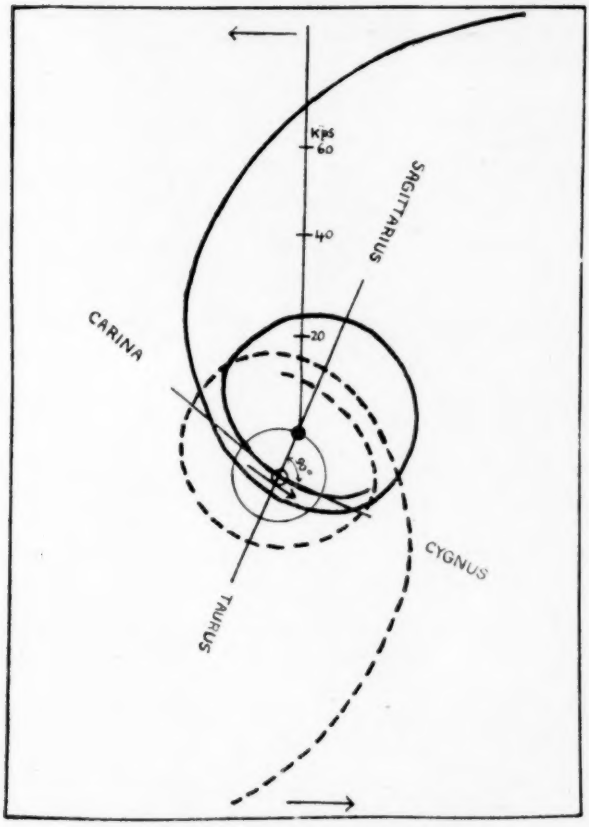


FIG. 8.—*Schematic representation of the Galaxy. The scale is for a solar distance from the galactic centre of 9·7 Kps.*

It will be seen that, for the solar position selected, the arms "lead" in the direction of the galactic rotation towards Cygnus, and that in the direction of Carina we are looking along the arm. The radio maximum and the Carina direction are indeed near the intersections between the two arms, though at a distance of about 20 Kps. the star-density at these points can hardly be observable. The overall dimensions of the central coils of the Galaxy have the enormous value of between 40 and 50 Kps.; but the stars in the outer regions would

be very sparsely distributed. The velocity in the circular orbit in this model is 260 km./sec., whilst the velocity in the solar neighbourhood was taken to be about 250 km./sec. Hence the rotation constants for the region of the Galaxy accessible to us would be little altered in our model.

The enormous dimensions derived for the galactic system are hardly surprising in view of the high mass of the Galaxy, which must cause it to increase in overall size more rapidly than a system of smaller mass. The ages derived for the spiral arms are not unreasonable, though perhaps somewhat high in view of current ideas as to the probable ages of supergiant stars. On the whole it seems possible to account for a number of features of our Galaxy—in particular the "leading" arm suggested by the radio observations—fairly well by means of our model system.

Harvard College Observatory,
Cambridge 38, *Mass.:*
1950 *December* 21.

CONTENTS

Printed in London by Taylor & Francis, Ltd., 1952 *January*